NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

DETECTION AND CLASSIFICATION OF BALEEN WHALE FORAGING CALLS COMBINING PATTERN RECOGNITION AND MACHINE LEARNING TECHNIQUES

by

Ho-Chun Huang

December 2016

Thesis Advisor: John Joseph
Co-Advisor: Tetyana Margolina

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE		*Form Approved OMB No. 0704–0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

1. AGENCY USE ONLY (*Leave blank*)	**2. REPORT DATE** December 2016	**3. REPORT TYPE AND DATES COVERED** Master's thesis	
4. TITLE AND SUBTITLE DETECTION AND CLASSIFICATION OF BALEEN WHALE FORAGING CALLS COMBINING PATTERN RECOGNITION AND MACHINE LEARNING TECHNIQUES		**5. FUNDING NUMBERS**	
6. AUTHOR(S) Ho-Chun Huang			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000		**8. PERFORMING ORGANIZATION REPORT NUMBER**	
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A		**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**	

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number ____N/A____.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE**

13. ABSTRACT (maximum 200 words)

A three-step approach has been developed for detecting and classifying the foraging calls of the blue whale, Balaenoptera musculus, and fin whale, Balaenoptera physalus, in passive acoustic recordings. This approach includes a pattern recognition algorithm to reduce the effects of ambient noise and to detect the foraging calls. The detected calls are then classified as blue whale D-calls or fin whale 40-Hz calls using a machine learning technique, a logistic regression classifier. These algorithms have been trained and evaluated using the Detection, Classification, Localization, and Density Estimation (DCLDE) annotated passive acoustic data, which were recorded off the Central and Southern California coast from 2009 to 2013. By using the cross-validation method and DCLDE scoring tool, this research shows high out-of-sample performance for these algorithms, namely 96% recall with 92% precision for pattern recognition and 96% accuracy for the logistic regression classifier. The result was published by the Institute of Electrical and Electronics Engineers (2016). The advantages of this automated approach over traditional manual methods are reproducibility, known performance, cost-efficiency, and automation. This approach has the potential to conquer the challenges of detecting and classifying the foraging calls, including the analysis of large acoustic data sets and real-time acoustic data processing.

14. SUBJECT TERMS blue whale, fin whale, foraging call, pattern recognition, machine learning, logistic regression classifier, detection, classification	**15. NUMBER OF PAGES** 89
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540–01-280-5500

Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

DETECTION AND CLASSIFICATION OF BALEEN WHALE FORAGING CALLS COMBINING PATTERN RECOGNITION AND MACHINE LEARNING TECHNIQUES

Ho-Chun Huang
Lieutenant Commander, Republic of China Navy
B.S., R.O.C. Naval Academy, 2003

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN PHYSICAL OCEANOGRAPHY

from the

**NAVAL POSTGRADUATE SCHOOL
December 2016**

Approved by: John Joseph
Thesis Advisor

Tetyana Margolina
Co-Advisor

Peter Chu
Chair, Department of Oceanography

iii

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

A three-step approach has been developed for detecting and classifying the foraging calls of the bl whale, *Balaenoptera musculus,* and fin whale, *Balaenoptera physalus,* in passive acoustic recordings. This approach includes a pattern recognition algorithm to reduce the effects of ambient noise and to detect the foraging calls. The detected calls are then classified as blue whale D-calls or fin whale 40-Hz calls using a machine learning technique, a logistic regression classifier. These algorithms have been trained and evaluated using the Detection, Classification, Localization, and Density Estimation (DCLDE) annotated passive acoustic data, which were recorded off the Central and Southern California coast from 2009 to 2013. By using the cross-validation method and DCLDE scoring tool, this research shows high out-of-sample performance for these algorithms, namely 96% recall with 92% precision for pattern recognition and 96% accuracy for the logistic regression classifier. The result was published by the Institute of Electrical and Electronics Engineers (2016). The advantages of this automated approach over traditional manual methods are reproducibility, known performance, cost-efficiency, and automation. This approach has the potential to conquer the challenges of detecting and classifying the foraging calls, including the analysis of large acoustic data sets and real-time acoustic data processing.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

x

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AM	amplitude modulated
DCLDE	Detection, Classification, Localization, and Density Estimation
FM	frequency modulated
FS	sampling rate
HARP	High-frequency Acoustic Recording Package
LTSA	Long-Term Spectral Average
NFFT	number of points used to form each fast Fourier transform
SIO	Scripps Institution of Oceanography
SNR	signal-to-noise ratio

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

The populations of two largest cetacean species, blue whales *Balaenoptera musculus* and fin whales *Balaenoptera physalus,* declined tremendously during the 20th century. More than 1.25 million blue and fin whales were killed by industrialized whaling from 1900 through 1989 (Rocha et al. 2014). Although commercial whaling was banned in 1982 under the International Whaling Commission's moratorium and statistical evidence has shown that the whale population has been increasing (Branch et al. 2004), they are still endangered.

Though blue and fin whales have been hunted for more than a century, little is known about their distribution, migration, and population. Surveys that aim to collect this type of information use data from catches, sightings, stranding episodes, discovery marks and recoveries, and acoustic recordings. While types of efforts differ substantially from area to area, sighting rate is the common qualitative measure to represent the status of the whale population (Branch et al. 2007). However, visual observations are limited due to such factors as weather, visual range, daylight, and whales' movements. Furthermore, whales spend more time underwater, and therefore the uncertainty of visual observations remains considerable.

Marine bioacoustic observation is another major approach to evaluate whale behavior, habitat, and population size. Blue and fin whales are known to produce various vocalizations. However, using whale vocalizations to estimate their populations is really a challenge, as is relating whale vocalizations with their behaviors. A whale is indeed present when a call is detected, but the lack of calls does not mean that whales are absent. They may just be quiet. Combining multiple call types into the analyses can diminish the bias due to the presence of quiet whales and help obtain more accurate estimation of the true presence (Širović et al. 2012). This research aims to develop an automated detector and classifier for blue and fin whales' foraging calls, which have the advantage of no gender exception. Currently, these calls are detected and classified by expert analysts, and thus limited documentation has been published. The data used in this research have been made available for the Detection, Classification, Location, and Density Estimation

1

(DCLDE) 2015 workshop held by Scripps Institution of Oceanography (SIO). This workshop provided annotations of the foraging calls, which have been considered the ground-truth for this research. The workshop also provided a scoring tool as a standardized way to estimate the performance of any detector or classifier of fin and blue whales foraging calls.

## A.	FIN WHALE VOCALIZATIONS

Fin whales produce two types of low-frequency sounds. The 20-Hz call, so called because the average frequency of the maximum intensity for the observed vocalizations in the early Atlantic recordings is centered around 20 Hz, is the most often reported fin whale sound worldwide (Watkins 1982; Edds 1988; Thompson et al. 1990; Watkins et al. 2000; Clark et al. 2002; Nieukirk et al. 2004; Širović et al. 2004; Castellote et al. 2012). However, only males have been found to produce 20-Hz calls, which might be a reproductive strategy (Croll et al. 2002). Another vocalization, the 40-Hz call, is produced by fin whales apparently in feeding contexts without gender exception (Watkins 1982). The frequency band of 40-Hz calls is generally 30–100 Hz, more often 40–75 Hz with down-sweeping in frequency (Širović et al. 2012). This call is more variable in character and only classified by expert analysts so far, but analyzing the call should enhance the accuracy of population estimates for fin whales.

## B.	BLUE WHALE VOCALIZATIONS

Several low-frequency sounds have been documented for blue whales. The best-described pulsed A-calls and tonal B-calls have well-defined frequency content, a long duration (15–20 seconds), and are produced in regular and repetitive sequences. Consequently, using a matched filter or spectrogram correlation method has proven effective for detection and identification of these calls. However, like the fin whale 20-Hz calls, these calls are only produced by males. Blue whales also produce highly variable D-calls related to feeding behavior (Wiggins et al. 2005; Oleson et al. 2007a), without gender exception. Characteristics of D-call include varying sweep rates of 25–90 Hz (Thompson et al. 1996; Oleson et al. 2007c) or 45–95 Hz (Madhusudhana et al. 2009), and duration of 1–4 seconds (Thompson et al. 1996; Oleson et al. 2007c; Madhusudhana

et al. 2009). Currently, this call is also classified by analysts through visual scanning of acoustic data. The fourth type of blue whale sound is a highly variable amplitude-modulated (AM) and frequency-modulated (FM) call, but the behavioral significance of the AM and FM call is unknown (Thode et al. 2000; Oleson et al. 2007a; Oleson et al. 2007c).

C.   VISUAL SCANNING PROTOCOL AND RESEARCH PROBLEMS

A traditional protocol (Oleson et al. 2007b) of visual scanning for the foraging calls applies the custom MATLAB-based program, Triton (Wiggins and Hildebrand 2007), to plot the Long-Term Spectral Average (LTSA) of calls, commonly using a bin size of 5-second-average and 1 Hz for a one-hour period with 0–250 Hz bandwidth. Analysts scan through the LTSA to locate candidates (i.e., potential signals of interest) of foraging calls, and then zoom in on the candidates to plot a short-term spectrogram to classify them as D-calls or 40-Hz calls. Parameters used for plotting the short-term spectrogram are chosen by analysts; this research uses a window of 9 seconds duration and 0–100 Hz bandwidth with 1 Hz and 0.09 seconds resolution. Classification of foraging calls not only depends on the call itself, but also on the contextual information analysts observed from the LTSA (i.e., presence of A-, B-, or 20-Hz calls in the analyzed period). Thus, the same data scanned by different analysts may have different results, and even the same analyst may have different results when scanning through the data again. This human factor in analyst performance implies there will always be a level of uncertainty in the final analysis produced by this method that is difficult to quantify. Furthermore, visual scanning of the data by human analysts is a labor-intensive process and, therefore, very costly approach.

Unlike most whale calls, there is no pattern analysts can observe from previous calls to predict subsequent blue and fin whale foraging calls. Blue whale D-calls typically have a distinctly broader bandwidth and longer duration than the fin whale 40-Hz calls, as shown in Figures 1B and 1C. However, it is difficult to classify the two call types in real data because of their similar downsweeping within a highly variable frequency range, as well as possible overlapping in call duration and frequency bandwidth. Blue and

fin whale foraging calls present a challenge for conventional detection techniques because of this higher variability as compared to other types of mysticete vocalizations, with perhaps the exception of humpback whale songs. Additionally, blue and fin whales are closely related species and their foraging calls sometimes have very similar characteristics, as shown in Figures 1D and 1E. Furthermore, the difficulty of identifying the calls is caused not only by features inherent to a specific call type but also by the interfering effects of ambient noise that may mask or distort the real calls in spectral displays.

Many techniques have been used for the automatic detection and classification of whales' vocalizations (Madhusudhana et al. 2010; Bahoura 2009; Bahoura et al. 2012; Helble et al. 2012; Shamir et al. 2014). However, the lack of ground truth samples for blue whale D-calls and fin whale 40-Hz calls had obstructed the development of their detector and classifier until SIO provided the annotated data for the DCLDE 2015 workshop. A machine learning technique known as deep learning has been used to classify the foraging calls since then (Karnowski et al. 2015). It shows the potential of machine learning with only in-sample performance, but indicates two issues associated with the method: 1) its inability to filter ambient noise sufficiently and 2) its requirement for an additional detector. This research uses pattern recognition to overcome both the issues and applies a more computationally efficient machine learning technique, logistic regression, to detect and classify the foraging calls (Huang et al. 2016). A cross-validation method is used to estimate the performance for this new approach.
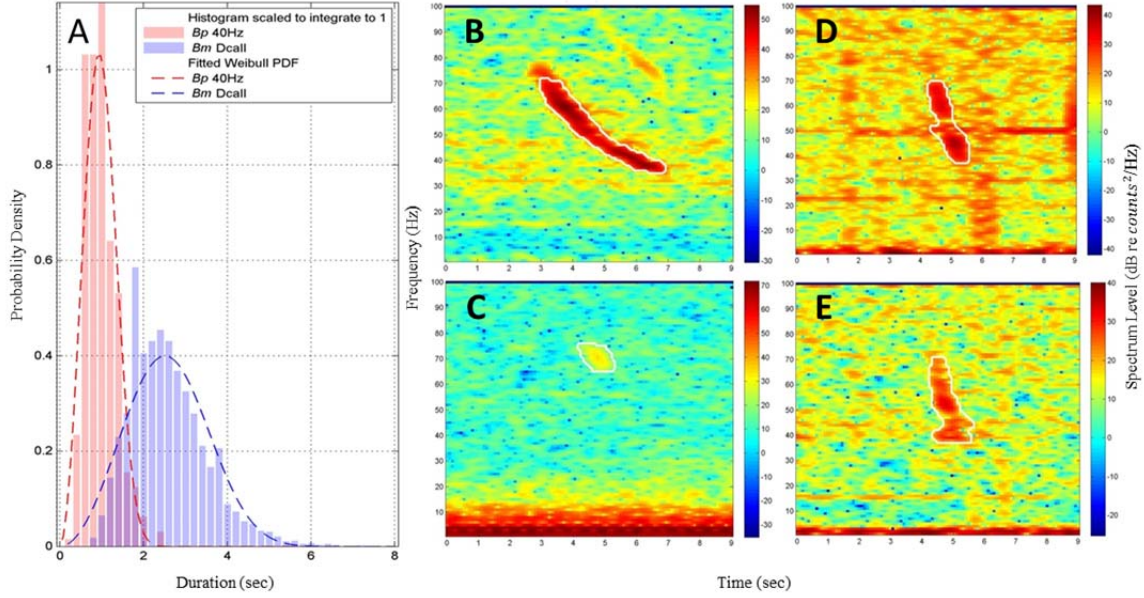
Image A shows the distribution of duration for 40-Hz calls and D-calls in the DCLDE 2015 dataset. Images B and C are spectrograms of a D-call (B) and a 40-Hz call (C). Images D and E are spectrograms of a short duration D-call (D) and a similar 40-Hz call (E).

Figure 1. Duration distribution and spectrograms of foraging calls.

D.   RESEARCH OBJECTIVES

This research aims to develop an automated detector and classifier for blue and fin whale foraging calls. The general concepts are shown in Figure 2. The detector uses data from acoustic recordings to generate spectrograms for designated periods and uses a pattern recognition algorithm (selection algorithm) to select the foraging call candidates. The purpose of the selection algorithm is to simulate the process of visually scanning the LTSA, which helps minimize the amount of data that requires further processing. Similar to the traditional protocol, the detector then plots a 9-second spectrogram for each candidate and uses another pattern recognition algorithm (detection algorithm) to de-noise the information and extract the call contour from the spectrogram. The logistic regression classifier (classification algorithm), a machine learning technique, is then applied to classify the contour as a D-call or a 40-Hz call. The hypothesis is that the logistic regression classifier should be able to more accurately classify D-calls and 40-Hz calls when applied to pre-processed images that contain only signals of interest with minimal ambient noise. The following are the three objectives of this research.

5

1.   Analyze the annotated data to discover patterns from foraging call spectrograms and develop a pattern recognition algorithm to de-noise the images and extract foraging calls' contours. The DCLDE scoring tool is then used to estimate the performance of the pattern recognition algorithm.

2.   Apply a logistic regression classifier to the extracted contours and use a cross-validation method to estimate the out-of-sample classification accuracy of the logistic regression prediction function.

3.   Develop another pattern recognition algorithm to select foraging call candidates from acoustic data and compare the results to the annotated data.
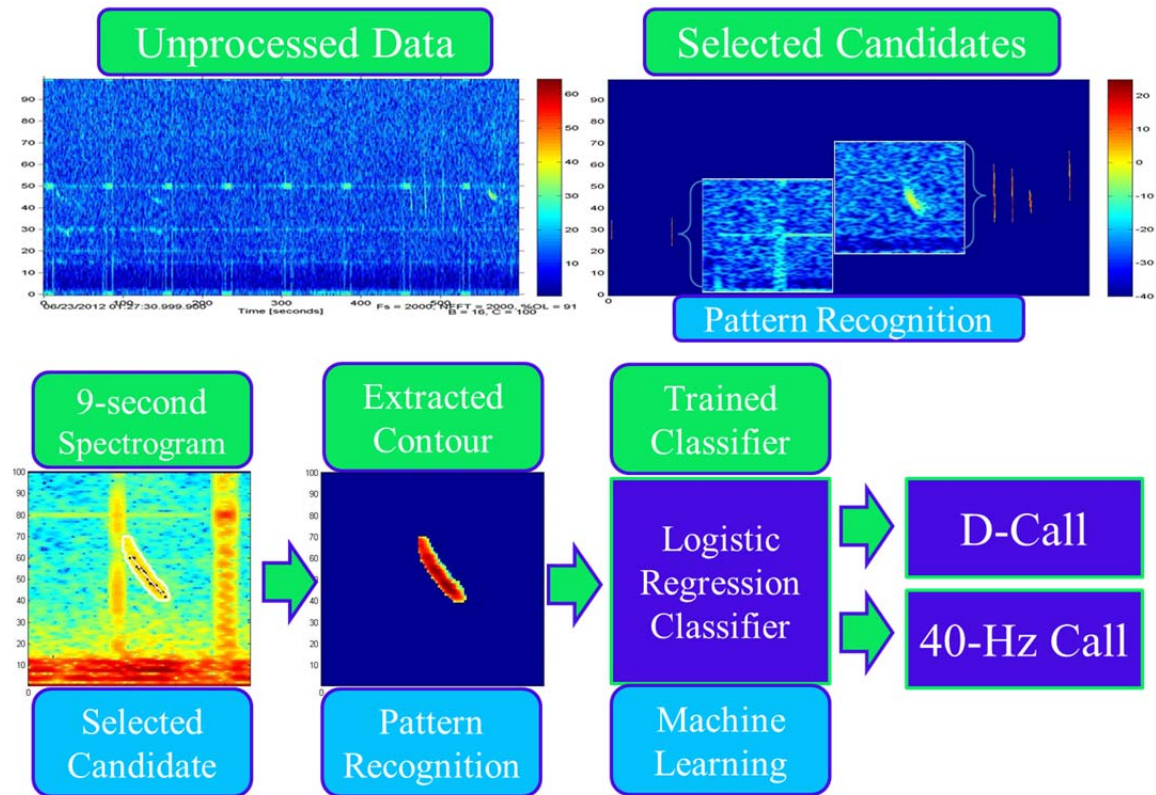
The top panels demonstrate the selection algorithm. The bottom charts demonstrate the tasks of the detection and classification algorithms.

Figure 2.   General concepts of the selection, detection, and selection algorithms.

E.   DATA

The DCLDE 2015 workshop provided two sets of recorded acoustic data, one focused on mysticete calls and the other focused on odontocete calls (SIO 2015). This

research uses only the mysticete dataset. The dataset has 19 uncompressed audio files originally recorded at high sampling rates and then decimated down to 1 and 1.6 kHz sampling frequency. These data were recorded between 2009 and 2013 from multiple High-frequency Acoustic Recording Packages (HARPs) (Wiggins and Hildebrand 2007), which were deployed in three different locations off the central and southern California coast as shown in Figure 3. The DCLDE 2015 workshop also provided an annotation file for blue whale D-calls and fin whale 40-Hz calls. The annotation provides information in six parameters: project name, site, species, start-time, end-time, and call type. It covers all seasons from the three locations, and contains 4,504 D-calls and 320 40-Hz calls used as ground-truth samples for this research.

DCLDE 2015 mysticetes data was recorded from three sites as shown by yellow pins. DCPP-A used a 320 kHz sampling rate at 65 m depth. DCPP-C used a 200 kHz sampling rate at 1000 m depth. CINMS-B used a 200 kHz sampling rate at 600 m depth.

Figure 3. Locations of DCLDE2015 low-frequency recording sites. Adapted from DCLDE (2015).

THIS PAGE INTENTIONALLY LEFT BLANK

II. METHODOLOGY

This research applies pattern recognition and logistic regression techniques to detect and classify blue and fin whale foraging calls. The whole process requires three algorithms for selecting candidates, i.e., potential foraging calls, detecting foraging calls from these candidates, and classifying foraging calls to species. Both selection and detection algorithms apply pattern recognition, and the classification algorithm applies a logistic regression approach. This research also uses the DCLDE scoring tool and cross-validation method to objectively estimate the performance of these algorithms.

The DCLDE annotation file is used to plot a 9-second spectrogram for each ground-truth call, providing 4,824 samples for this research. The cross-validation method requires these samples to be randomly assigned into three subsets: training (60%), testing (20%), and validation (20%). Thus, each subset still contains characteristics of the different physical locations and seasons. The training subset is used for extracting patterns of the foraging calls and to build the detection and classification algorithms. The testing subset is used for estimating the initial performances of both algorithms. These initial performances provide clues for tuning both algorithms. The training and testing subsets can be used as many times as necessary to develop the optimal algorithms; however, the validation subset is used only once to assess the final "out-of-sample" performance of these algorithms.

To clarify, the training and testing subsets are merged into an in-sample subset for developing the final prediction function. Applying this prediction function to the in-sample subset should provide very good performance, the so called in-sample performance, since the prediction function has already been fitted to this subset. On the other hand, applying the prediction function to the validation subset, the so called out-of-sample subset, that has not been used to develop the prediction function, provides out-of-sample performance. High out-of-sample performance indicates that the prediction function generalizes well.

This chapter first describes spectrogram parameters and then covers three algorithms following the sequence in which they were developed. Since features of the foraging calls are learned from the DCLDE annotated calls, and then are applied to select candidates from the original acoustic data, the detection algorithm was developed first. The detection algorithm applies pattern recognition to de-noise and extract call contours from the annotated 9-second spectrograms. Its output is the input to the classification algorithm, and its rules and thresholds are the cornerstone for the selection algorithm. The classification algorithm applies logistic regression to classify the extracted contours as blue whale D-calls or fin whale 40-Hz calls. The selection algorithm also applies a pattern recognition scheme but aims to scan the original acoustic data to select potential foraging calls. This algorithm uses most of the call-oriented rules from the detection algorithm. However, it also needs noise-oriented rules, which require additional observations from the acoustic dataset. Thus, the selection and detection algorithms are developed and discussed individually. This chapter covers details of the development of the three algorithms. The results show that the detection and classification algorithms are well-designed but the selection algorithm requires further development and assessment.

A.   SPECTROGRAM DEFINITION

As noted earlier, the 9-second spectrograms of ground-truth foraging calls are the cornerstones for this research. Analysts have the freedom to adjust any parameter in Triton for plotting a spectrogram; however, this research requires a set of fixed parameters to create the spectrograms so that the algorithms compare "apples to apples" and not "apples to oranges." The parameters used in this research are adapted from protocols established by SIO and the Ocean Acoustic Laboratory of the Naval Postgraduate School (Oleson 2007a; Margolina 2010; Širović 2011).

The temporal center of a 9-second spectrogram is the average of each DCLDE annotated call start time and end time. Parameters include a 10-second duration, a 0–100 Hz frequency band, the number of points used to form each fast Fourier transform (NFFT) equal to the sampling rate (FS), 91% overlap of FFT segments, 100% contrast setting in Triton, 16% brightness setting in Triton, and the blue-red "jet" color-map used

in MATLAB. With these parameters, Triton plots a 9.09-second spectrogram from 0 Hz to 100 Hz with a resolution of 1 Hz and 0.09 seconds. Each image is viewed as a 100×101 matrix resulting in 10,100 elements. The large values of NFFT and overlap are adapted from the original protocols used by Oleson, Margolina, and Širović to provide better resolution for defining features with pattern recognition methods. Theoretically, a finer resolution demonstrates more detail of an object. However, the tradeoff between temporal and frequency resolutions requires a balance to compromise them. The minimum bandwidth and duration for a foraging call is approximately 5 Hz and 0.5 seconds, respectively. These settings provide a frequency resolution of 1 Hz and time resolution of 0.09 seconds. Thus, both frequency and temporal features of a foraging call can be represented by at least five elements of the spectral output matrix. The longer duration spectrograms used for the selection algorithm apply the same parameters except duration. Thus, all spectrograms have the same resolution and the features of the foraging calls are the same as well.

B. PATTERN RECOGNITION FOR DE-NOISING AND CONTOUR EXTRACTION

The pattern recognition technique follows the mathematical and technical aspects of human perception, or the knowledge of the expert analyst about foraging calls. The detection algorithm, which uses pattern recognition for de-noising and contour extraction, can be viewed as a mapping of a foraging call from its 9-second spectrogram into a highly reduced-noise space. Since the output of this algorithm are call contours in an otherwise noise-free space, the logistic regression classifier is able to classify the contours more accurately than from the original spectrograms.

A spectrogram is a three-dimensional image containing time, frequency, and intensity information. The characteristics of the foraging calls make it possible to discover patterns in their spectrograms. The patterns of foraging calls in the 9-second spectrograms are translated into rules and thresholds in the detection algorithm, which applies computer vision tools for image processing. The fundamental rules are based on expert knowledge about the foraging call features such as frequency downsweeping between 20 and 100 Hz and short duration (less than 9 seconds). However, conquering

the task requires more precise rules and thresholds, which can be observed from the DCLDE annotated calls. The goal of the detection algorithm is to filter as much noise as possible and to precisely extract the call contours. The detection algorithm has four major rules, as shown in Figure 4, and the details are described in the following sub-sections.
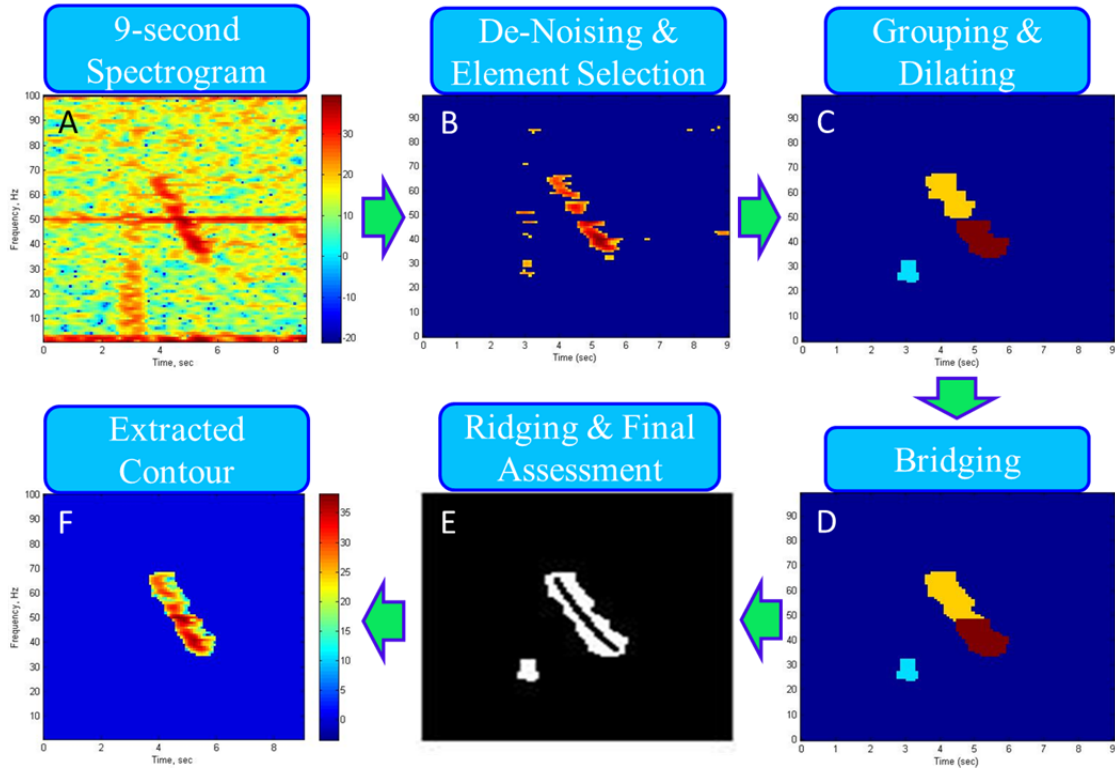
Image A is a 9-second spectrogram plotting from annotation data. Image B is 4% of the elements selected by the first rule. Image C shows three contours kept by the second rule. Image D shows two contours remaining after bridging. The black line of the white contour in image E is the frequency downsweeping feature detected by the ridging rule. Image F shows the final contour extracted by the detection algorithm.

Figure 4. Flow chart of the detection algorithm.

1.   De-Noising and Elements Selecting

The detection algorithm applies a rule that filters most noise and keeps only 4% of the highest intensity elements between 25 and 90 Hz. This rule considers all elements equally to determine signal and ambient noise simultaneously for determining the 4% threshold, which is similar to the concept of choosing a detection threshold for a sonar system by analyzing the signal-to-noise ratio (SNR). This rule mitigates the noise effects

by applying three filters, and assumes that only 4% of the elements are part of a foraging call and the rest are useless. Keeping only 4% elements is an empirical rule that balances the tradeoff between recall and precision. More details of how to determine the threshold are discussed in the next chapter.

The first filter is a frequency filter. Initially, the band 20–90 Hz was applied to the training subset to establish the algorithm. In the training subset, only 1.73% (50/2895) annotated calls were found to contain information above 90 Hz, and no calls were found to have information below 20 Hz. After the algorithm was tuned by the initial performance estimation, four frequency bands (20–90 Hz, 25–90 Hz, 20–95 Hz, and 25–95 Hz) were tested to define the bandwidth limits applied in the final detection algorithm, and the band of 25–90 Hz was found to be optimal. The lower limit can filter out the fin whale 20-Hz calls, which are often present at the same time as foraging calls. The upper limit can filter out all sound sources above 90 Hz.

The objective of the second filter targets tonal and broadband noise present in the DCLDE dataset. Tonal noise is defined as high intensity, long duration, and narrow band noise. It typically appears as a horizontal red line in the spectrograms, as shown in Figures 5A and 5B. Tonal noise in a 9-second spectrogram can be from instrument noise, shipping noise, or even a blue whale A-call or B-call. Broadband noise, on the other hand, appears as an area of relatively high intensity and short duration covering a wide frequency band. It typically looks like a vertical red line in the spectrograms, as shown in Figure 5C. Sources of broadband noise can be instrument self-noise, fishing equipment sound, air gun sound, and even other unidentified marine mammal vocalizations. Since these noises can mask the foraging calls while the algorithm is selecting 4% of the highest intensity elements, these need to be filtered before the selection task is done.
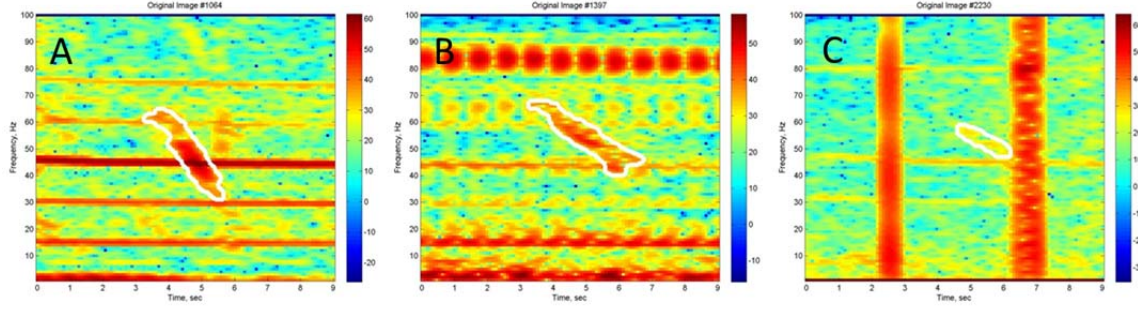
Images A, B, and C are the spectrograms of annotated calls #1064, #1397, and #2230 in the in-sample dataset (merged training and testing subset together). Without filtering out the noise, it is impossible to extract calls' contours (white outline) from the spectrograms.

Figure 5. Examples of tonal and broadband noise.

The filter sets intensity values for out of band and negative intensity elements equal to zero, and then calculates the intensity threshold (I_n):

$$I_n = \overline{I} + \left(I_{max} - \overline{I}\right) \times a$$

(1)

where $\overline{I}$ is the mean intensity, I_{max} is the maximum intensity, a = 0.1 for the detection algorithm, and a = 0.4 for the selection algorithm. If more than 95% of the elements of any row or column have an intensity larger than I_n, the filter will set the intensity value of this row or column equal to zero. After applying the first filter, the algorithm starts selecting elements from the spectrogram.

The element selecting loop requires an initial intensity threshold (I_i) to begin with:

$$I_i = I_{max} - \frac{I_{max} \div \overline{I}}{12} \times \left(I_{max} - \overline{I}\right)$$

(2)

where a higher initial threshold is used to minimize the computational effort. The assumptions are that $\overline{I}$ decreases with low ambient noise and I_{max} increases due to a loud foraging call. Thus, I_i increases as $\overline{I}$ is decreasing or I_{max} is increasing. Elements with

14

intensity lower than this threshold are set equal to zero. If the amount of selected elements is less than 4% of in-band positive intensity elements, this threshold decreases by $0.001 \times \left(I_{max} - \overline{I} \right)$ until at least 4% of the elements pass it.

The third filter, a de-noise function, is applied while selecting elements. This filter targets smaller ambient noise elements covering not entire columns or rows but still a majority of them. If a row has more than 35 nonzero elements (35% of the elements of a row) or a column has more than 39 nonzero elements (60% of the elements of a column) after element selection, the de-noise function sets the values of this row or column equal to zero and selects additional elements until meeting the 4% threshold again.

The 4% threshold is a dynamic number because the denominator is the number of positive intensity elements within the 25–90 Hz frequency band after filtering the noisy columns and rows. Though the contours of some calls are not preserved exactly, the overall performance is not affected by this imperfection. The elements selected by this rule can be viewed as the pieces of a foraging call puzzle. The following rules aim to put these pieces together.

The threshold of 4% of the highest intensity elements was defined empirically by comparing the results of using different thresholds. There is always a tradeoff between contour-preservation and noise-filtering. Furthermore, it is not easy to use a fixed threshold to preserve contours of different call types since D-calls have longer duration on average and a broader bandwidth than 40-Hz calls, which means more elements are frequently needed to represent D-call contours. However, more ambient noises elements would be included once the element percentage is increased, especially in the case of 40-Hz calls. Consequently, some of the D-call contours include only a portion of the call elements and some ambient noise elements are kept after applying the 4% threshold.

2. Grouping and Dilating

The rule of grouping and dilating filters out small random noise features and creates initial contours. Approximately 250 elements are selected from the first rule. A majority of them are parts of a foraging call. This rule groups elements that connect to

each other, and keeps only the groups that have more than six elements. The remaining groups are dilated by a 3×3 matrix. The function of dilation is shown in Figure 6. Elements are grouped again and only groups with more than 45 elements are kept.

Different sizes and combinations of the dilation matrix (as shown in the appendix) as well as dilation number of times have been tested for determining the final thresholds for this rule. The effects of using bigger dilation matrices or repeating the dilation process multiple times are similar but not the same. Basically, dilation means the algorithm uses more elements than the number selected in the first rule. However, increasing the 4% threshold in the first rule is not as beneficial as dilation. Dilation is similar to enlarging the size of each group but not generating useless groups, which is the effect of selecting more than enough elements.
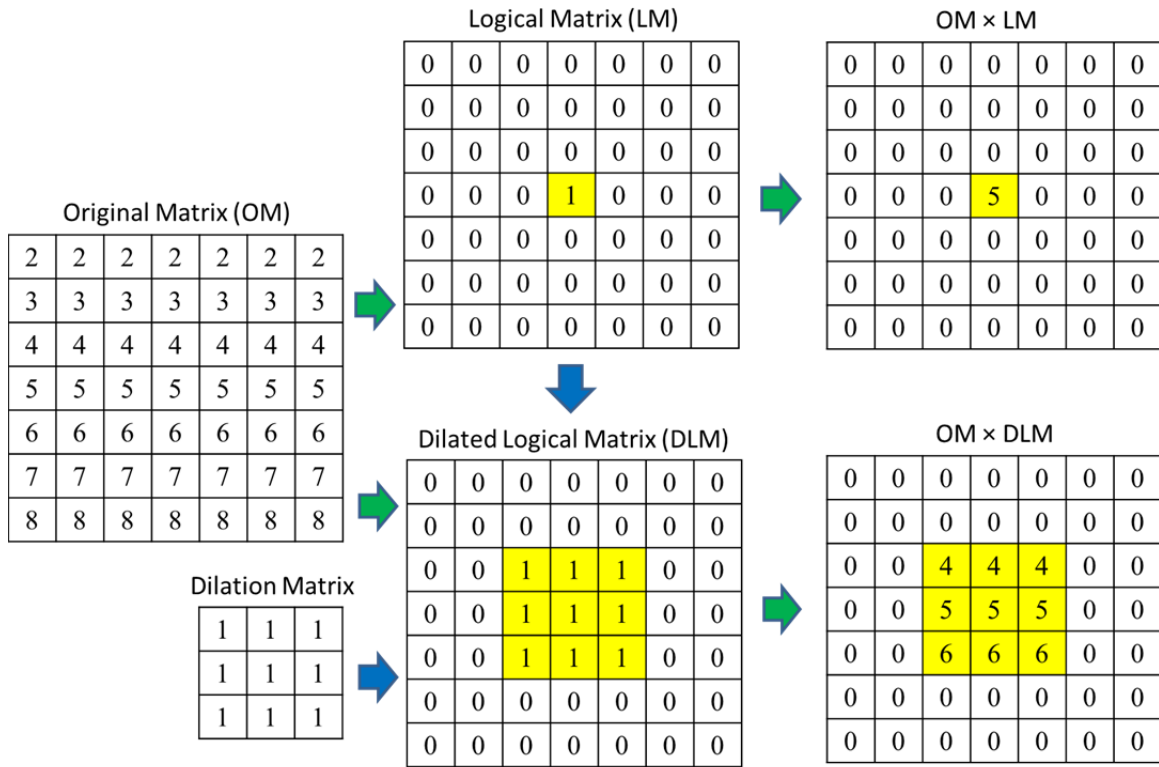
The original matrix can be viewed as an image formed of 7×7 pixels. The logical matrix controls how many pixels are used to represent the image. Dilation increases the size of logical matrix. Thus, more pixels can be used to represent the image.

Figure 6. Function of dilation.

16

The contours are smoother and bigger after dilation; however, over-dilation may mask the characteristics of a foraging call by creating a contour that is too smooth or too large. Furthermore, dilation can sometimes generate a noise contour with features similar to those of a call contour. Additionally, different dilation matrices generate different shapes and consequently provide different results. Using the final dilation matrix only once is based on the estimates of the overall performance for the detection algorithm on the training subset. After dilation, a spectrogram may have multiple contours and the next rule, bridging, connects most contours that belong to the same foraging call.

3.   Bridging

The bridging rule connects broken parts of the spectrogram of a foraging call and fills in empty spaces within a contour. A call may appear to be separated into several parts in the spectrogram image because of removing broadband and tonal noise. It may also just have multiple parts due to random discontinuity in the spectrogram itself. Gaps of broken parts can be defined in the display by two factors: distance and direction. Consequently, the thresholds of this rule are set as angle and length of a bridge, which are required for connecting these broken parts. Since the downsweeping slopes of the foraging calls and the width or height of the gaps vary, it is not easy to determine an adequate bridge to connect parts of a single call but not to connect call and noise. By evaluating the training subset, three bridges are used in this rule. All three bridges have the same length of five elements but use different angles of 20, 45, and 70 degrees (calculated by number of bins), respectively. The lengths of the three bridges are fixed since the gap is usually created by the removal of a tonal or broadband noise feature, and these noises have similar features. However, the frequency downsweeping slope for foraging calls is very different. Thus, different angles and multiple bridges are required.

Bridging not only connects corresponding parts of the displayed call, but also makes the contour even smoother. Discontinuities in sound are rare and the side effect of this rule is the possibility of producing artificial contours, which may cause a noise feature to be detected as a false positive or connect a noise feature to a foraging call. Consequently, this rule is only useful for a specific condition—where many tonal and

broadband noise features are present in the spectrogram, which is the case of most DCLDE data, as shown in Figure 7. The dynamic feature of ambient noise implies the thresholds of this rule should be modified when the algorithm is applied to other datasets. It also implies that the performance of the detection algorithm can be improved by mitigating the effects of instrument noise on detection. The effects of using different lengths of bridges are demonstrated in Figure 8. The best result of this rule is that there is only one contour for each 9-second spectrogram and this contour represents the annotated foraging call. However, some noise contours meet all the criteria of the previous rules and are still present. The next rule examines the detail features of each contour to filter out the noise contours.
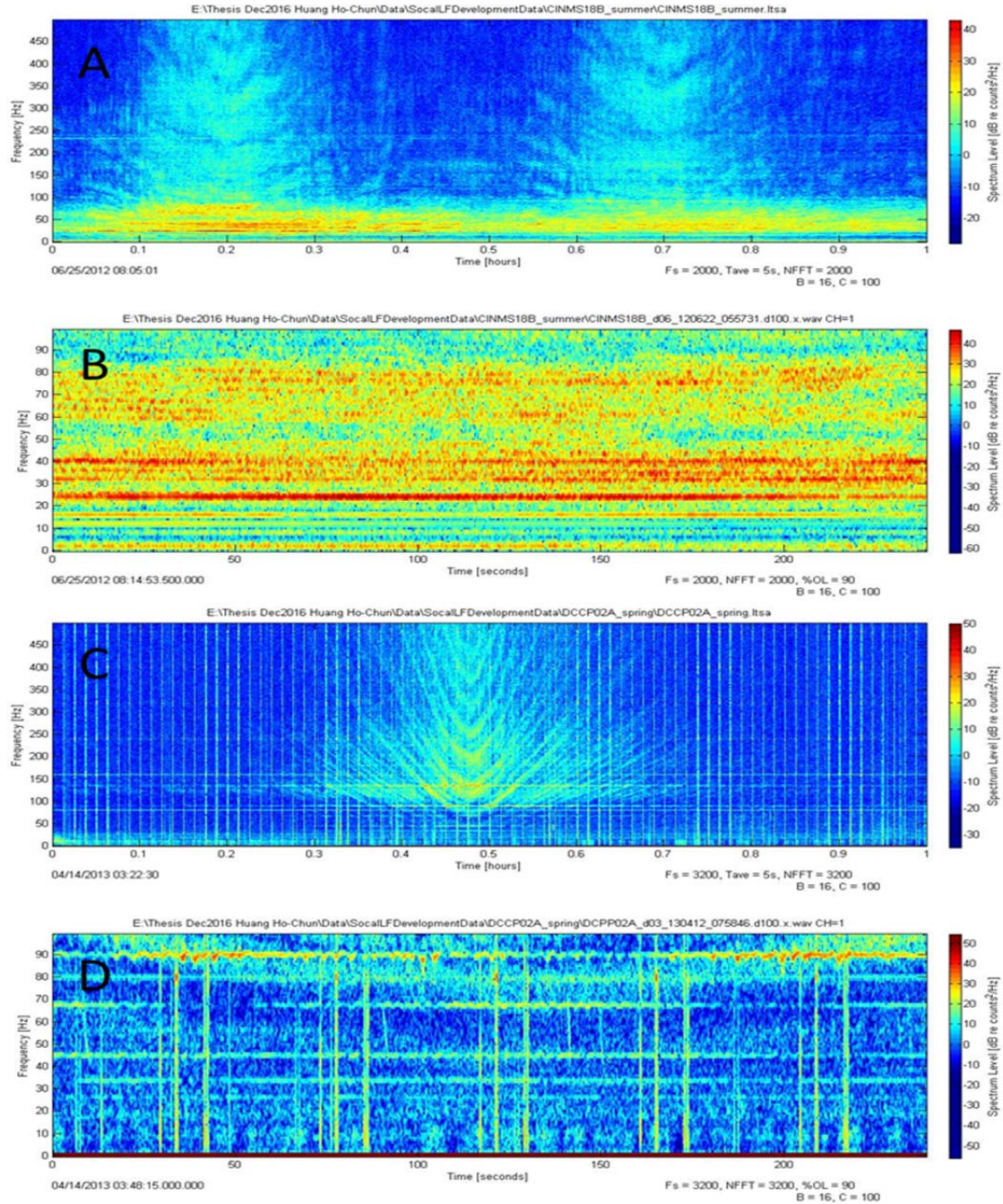
Image A is a 1-hour LTSA starting on 06/25/2012 08:05:01. Image B is a 4-minute spectrogram starting on 06/25/2012 08:14:53. The tonal noise, represented by red horizontal lines in both A and B, is most likely the shipping noise. Image C is a 1-hour LTSA start from 04/14/2013 03:22:30. Image D is a 4-minute spectrogram start from 04/14/2012 03:48:15. The broadband noise, represented by yellow vertical lines in both C and D, is most likely the instrument noise. The tonal noise, the yellow and orange horizontal lines in C and D, is most likely the instrument noise enhanced by the shipping noise.

Figure 7.   Source of the tonal and broadband noise.
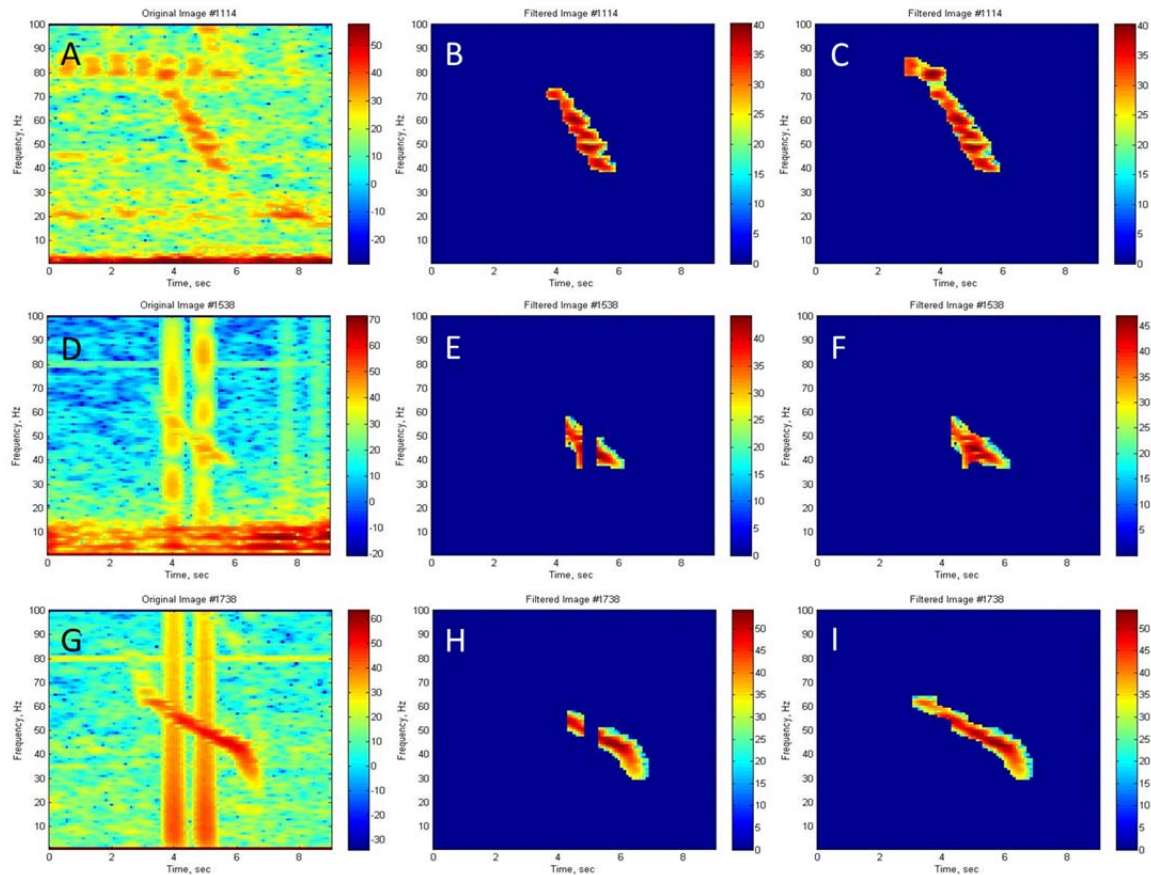
The left column is the original spectrograms of #1114 (A), #1538 (D), and #1738 (G) annotated calls in the training subset. Each row shows the results of using different bridge length. Image B applies a 1-element-bridge and image C applies a 3-element bridge, which connects the call contour to a noise feature that is most likely the end of a blue whale A-call. Image E applies a 3-element- and image F applies a 5-element bridge, which is able to connect the broken parts of image E. Image H applies a 5-element bridge and image I applies a 7-element bridge, which is able to connect the broken parts of image H.

Figure 8. Effects of different bridge lengths.

4.   Ridging and Final Assessment

The rule of ridging and final assessment determines the contours as a foraging call or a noise feature. Each 9-second spectrogram may present multiple contours after applying previous rules. Some of the contours are ambient noise features, which unexpectedly pass the thresholds of all the rules applied thus far. This rule applies several specific criteria to examine each contour. To begin with, a re-denoise function is used to further remove the noise "tail" or "head" from the contour, as shown in Figure 9. Though

20

the first application of the de-noise rule is able to remove most of the tonal noise, some contours may still have tonal noise features attached because these features do not meet the threshold of previous filters. The re-denoise function focuses on noise features with a narrow and fixed bandwidth and relatively long duration. The details of this function are described in the next paragraph.
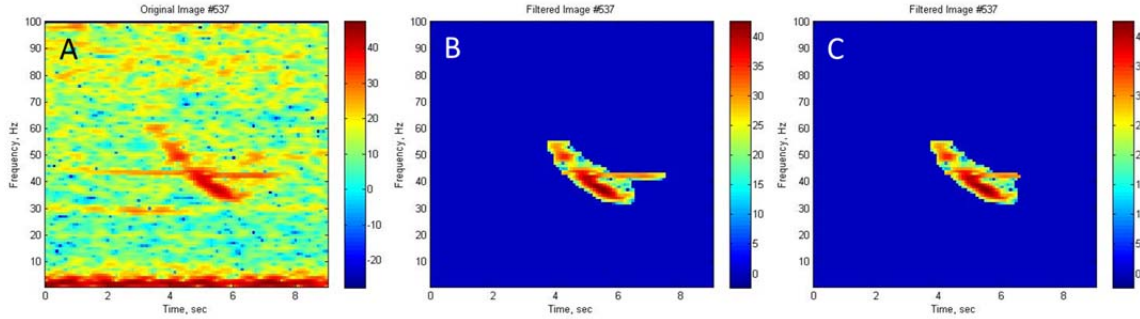
Image A is the original 9-second spectrogram of annotated call #537 in the in-sample dataset. Image B shows the extracted call contour before applying the re-denoise function. A noise feature attaches to the call contour like its tail. Image C shows the contour after applying the re-denoise function, which removes the noise tail to better estimate call duration.

Figure 9. Removing noise tail by the re-denoise function.

The task of the re-denoise function is to find and remove residual noise features of a contour. Each column of a tonal noise feature usually has a uniform size of less than 5 elements. Each column of a foraging call usually has varying size of at least 10 elements. This function calculates the bandwidth of individual temporal column and uses the minimum non-zero bandwidth value to locate the noise feature (e.g., the minimum bandwidth value of the contour in Figure 9B is 126 of the noise feature that has a band of 41–43 Hz). If more than 10 temporal bins have the same minimum bandwidth value, the re-denoise function will set the intensity value of these bins equal to zero, as shown in Figure 9C. The assumption of this function is that the duration of the tonal noise should be relatively long (≥ 0.9 seconds); otherwise the algorithm can just ignore it. Removing noise is critical for the following steps, which check contour duration, bandwidth, ratio of bandwidth to duration, and frequency downsweeping characteristics.

Frequency downsweeping is a critical characteristic of foraging calls. The downsweeping feature can be viewed as a negative slope for the ridge of a foraging call. The ridge of a call is formed by the highest intensity elements along each temporal column. The frequency of these elements should decrease in time if the contour is a foraging call. Two nearby elements in the ridge represent a segment. A ridge needs to have negative slope on average and at least three frequency downsweeping segments to pass the criteria. Additionally, the frequency difference of a segment must be less than 7 Hz since a dramatic frequency discontinuity is most likely due to two sound sources accidentally connected by the bridges.

After checking the downsweeping feature, the algorithm uses two sets of criteria to further examine the contours. The first set of criteria ensures that bandwidth is larger or equal to 9 Hz, duration is larger or equal to 0.45 seconds, ratio of bandwidth to duration is larger or equal to 0.4 (frequency bins/temporal bins), maximum frequency is larger than 35 Hz, and minimum frequency is smaller than 75 Hz. The second set ensures duration is larger or equal to 2.25 seconds, ratio of bandwidth to duration is larger or equal to 0.3, maximum frequency is larger than 35 Hz, and minimum frequency is smaller than 75 Hz.

Initially, the first set is designed for all blue whale D-calls and fin whale 40-Hz calls. However, some D-calls cannot meet the first set of criteria because of their long duration and gentle slope of frequency downsweeping, as shown in Figure 10. Thus, the second set is required. Both sets include the threshold that maximum frequency should be larger than 35 Hz. This threshold aims to filter the fin whale 20-Hz calls, which commonly accompany the foraging calls and are sometimes have frequency content in a band higher than 20 Hz. There is little tradeoff for this threshold since at least one D-call in the training subset is known not to pass this threshold.

Both sets also apply the threshold that minimum frequency should be smaller than 75 Hz. This threshold filters sound sources that are completely present above 75 Hz. Blue whale A-calls, for example, are one of these sound sources. A strong harmonic of A-calls starts at around 80–90 Hz with gentle slope of frequency downsweeping. Without this threshold, a part of an A-call present in a 9-seconod spectrogram (as shown in Figure 8A)

may be mistakenly extracted by the detection algorithm. It is assumed that there is no tradeoff for this threshold since all annotated foraging calls in the training subset pass this threshold.
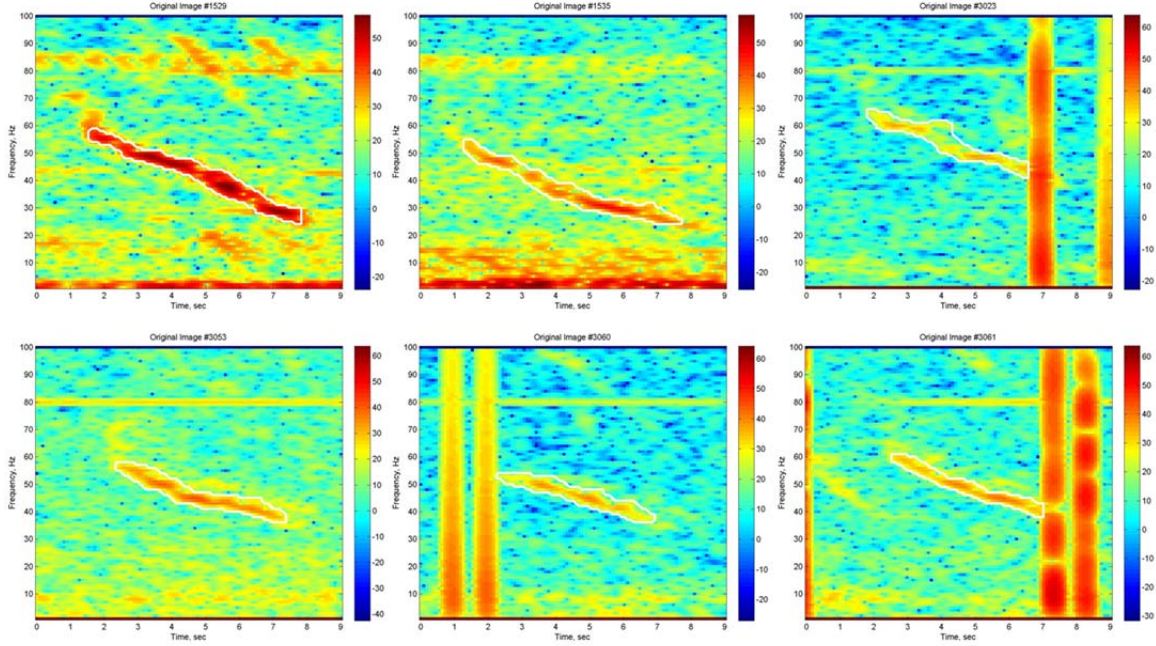
These are original 9-second spectrograms of annotated calls #1529, #1535, #3023, #3053, #3060, and #3061 in the in-sample dataset (merged training and testing subset together). Their ratios of bandwidth to duration are less than 0.4. Thus, they require a second set of criteria.

Figure 10. Long duration blue whale D-calls.

Final thresholds for all rules are listed in Table 1. The detection algorithm applies these thresholds and produces two files. One file contains only temporal information of all contours and uses the DCLDE annotation format so the scoring tool can be applied to compare the detector output to the DCLDE ground truth. The other file is the input for the logistic regression classifier and contains all information of both contours and their original spectrograms. Not all extracted contours can be put in this file but only contours centering in the original 9-second spectrograms. It is assumed that a contour located in different position in the spectrogram is most likely a noise feature or a portion of another foraging call since a 9-second spectrogram is centered on an annotated foraging call. It is better to remove this contour since it may mislead the classification algorithm.

23

Table 1.   Thresholds of the detection algorithm

Threshold	Value	Note
Percentage of element selection	4%	
Frequency celling	90 Hz	
Frequency bottom	25 Hz	
I_n	Equation 1	
I_i	Equation 2	
De-noise function for tonal noise	Any row keeps more than 35% of its elements	After 4% of the elements are selected
De-noise function for broad band noise	Any column keeps more than 60% of its elements	After 4% of the elements are selected
First grouping size	larger or equal to 6 elements	Before dilation
Second grouping size	larger or equal to 45 elements	After dilation
Dilation matrix	$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$	
Dilation number of time	1	Three is the maximum for the detection algorithm
Amount of bridges	3 bridges	
Length of bridges	5 elements	Same length for all bridges
Angle of bridges	20, 45, and 70 degrees	Calculated by number of bins
Re-denoise length	larger or equal to 10 columns	Same minimum frequency values for 10 columns
Frequency difference for each ridge segment	less or equal to 7 Hz	Frequency difference of a ridge segment ≤ 7 Hz
Qualify ridge	3 qualified segments	
First set of criteria	bandwidth ≥ 9 Hz, duration ≥0.45 seconds, bandwidth/duration ≥ 0.4, maximum frequency ≥ 35 Hz, minimum frequency ≤ 75 Hz	Ratio of bandwidth to duration is calculated by number of bins
Second set of criteria	duration ≥ 2.25 seconds, bandwidth/duration ≥ 0.3, maximum frequency ≥ 35 Hz, minimum frequency ≤ 75 Hz	Ratio of bandwidth to duration is calculated by number of bins

All thresholds can be changed when adapting for different environments and sensors.

C. LOGISTIC REGRESSION CLASSIFIER

Machine learning is a method of data analysis using computer algorithms for simulating the human learning processes to make these algorithms become intelligent systems. These algorithms have the ability to learn from data. The self-learning ability of machine learning is crucial for this research since there are limited descriptions based on human-expert analysis that define the differences between blue whale D-calls and fin whale 40-Hz calls.

This research uses a logistic regression classifier as the machine learning classification algorithm. The logistic regression classifier is one of many machine learning algorithms readily available through the Statistics and Machine Learning Toolbox in Matlab. It is able to efficiently classify a matrix of data similar to the spectrograms in this research. The cross-validation method is used to train, tune, and evaluate the classification algorithm. This algorithm was modified based on concepts learned from the on-line course *Machine Learning*, created by Stanford University and taught by Professor Andrew Ng (available online at https://www.coursera.org/learn/machine-learning). General concepts of the logistic regression and the cross-validation steps are introduced in the first subsection. The symbols and equations used in this research were adapted from Yaser et al. (2012). More detail can be found in Chapter 3.3 of this book or on the Stanford website of the on-line course. The second subsection explains why the logistic regression requires cross-validation and pattern recognition.

1. Logistic Regression Concepts

Logistic regression classifiers develop a prediction function to classify designated objects, e.g., blue whale D-calls and fin whale 40-Hz calls in this research. These calls are represented by their spectrograms, which are matrices containing information of time, frequency, and intensity. To apply the readily available Matlab algorithms, these matrices are converted to vectors. The coordinates of a spectrogram (i.e., time on the x-axis and frequency on the y-axis) are converted to the order of an input vector. The intensity information becomes the value of each element in the vector. Consequently, an image

becomes a vector with 10,100 elements. Theoretically, a function f is able to exactly classify two calls:

$$P(\mathrm{y}|X) = \begin{cases} f(X) & \text{for } y = 1 \\ 1 - f(X) & \text{for } y = -1 \end{cases}$$

(3)

where X is an input vector, and $y = 1$ for D-calls, and $y = -1$ for 40-Hz calls. Due to f being an unknown function, logistic regression develops a prediction function h as close as possible to the function f:

$$P(\mathrm{y}|X) = \begin{cases} h(X) & \text{for } y = 1 \\ 1 - h(X) & \text{for } y = -1 \end{cases}$$

(4)

The prediction function applies a sigmoid threshold θ for classifying designated objects:

$$h(X) = \theta(s), \ \theta(s) = \frac{1}{1 + e^{-s}}, \ s = W^T X$$

(5)

where s is the modified input, W is a weight vector determined by the *perceptron learning algorithm*. This weight vector reflects that different coordinates of X have different importance in the classification decision. More detail can be found in Chapter 1.1 of the book *Learning from data* (Yaser, Malik, and Lin 2012). Using a sigmoid threshold allows the model to smoothly restrict the output to the probability range, as shown in Figure 11A, and is substituted for the function h:

$$P(\mathrm{y}|X) = \theta(yW^T X)$$

(6)

The hypothesis, for which the input objects are either D-calls or 40-Hz calls, can be more efficiently represented as:

$$\prod_{n=1}^{N} P(\mathbf{y}_n \mid X_n) \tag{7}$$

where N is the total number of input vectors. Since each input vector is independently generating a probability, the probability of getting all the y_n in the data from the corresponding X_n would be equal to (7). The maximum likelihood method helps the prediction function h maximize the overall probability. This task is equivalent to minimizing the product of:

$$-\frac{1}{N}\ln\left(\prod_{n=1}^{N} P(\mathbf{y}_n \mid X_n)\right) = \frac{1}{N}\sum_{n=1}^{N}\ln\left(\frac{1}{P(\mathbf{y}_n \mid X_n)}\right) = \frac{1}{N}\sum_{n=1}^{N}\ln\left(\frac{1}{\theta\left(y_n W^T X_n\right)}\right) \tag{8}$$

It can be viewed as the logistic regression model is minimizing an error measure, which is the in-sample error $E_{in}(W)$:

$$E_{in}(W) = \frac{1}{N}\sum_{n=1}^{N}\ln\left(1 + e^{-y_n W^T X_n}\right) \tag{9}$$

The approach of minimizing the error applies a gradient descent technique. Gradient descent is used for minimizing a twice-differentiable function, which is similar to $E_{in}(W)$. The physical analogy of gradient descent is a ball rolling down a bowl shaped hill, as shown in Figure 11B. The bottom point of the hill represents where the minimum in-sample error is. The task of gradient descent is to move the ball as close to the bottom point as possible. The change of error $\Delta E_{in}(W)$ is *the* error after each iteration subtracts the original error:

$$\Delta E_{in} = E_{in}\left(W(0) + \eta\hat{v}\right) - E_{in}\left(W(0)\right) = \eta \nabla E_{in}\left(W(0)\right)^T \hat{v} + O\left(\eta^2\right) \geq -\eta\left\|\nabla E_{in}\left(W(0)\right)\right\| \tag{10}$$

where η is the step size of each iteration, and $\hat{v}$ is a unit vector of direction for the iteration. The new weight vector after the first iteration is $W(0) + \eta\hat{v}$. A misclassified call

contributes more to the $\nabla Ein(W)$, which is the gradient (g), than a correctly classified one. Thus, the gradient descent technique aims to minimize the error, i.e., products of:

$$g_t = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n X_n}{1+e^{y_n W_{(t)}^T X_n}}$$

(11)

using $-g_t = v_t$ and updating the weights ($W(t+1)=W(t)+\eta v_t$) until the designated time (e.g., 100 iterations in this research) to stop is reached. The final weight vectors are shown in Figure 11C and D.
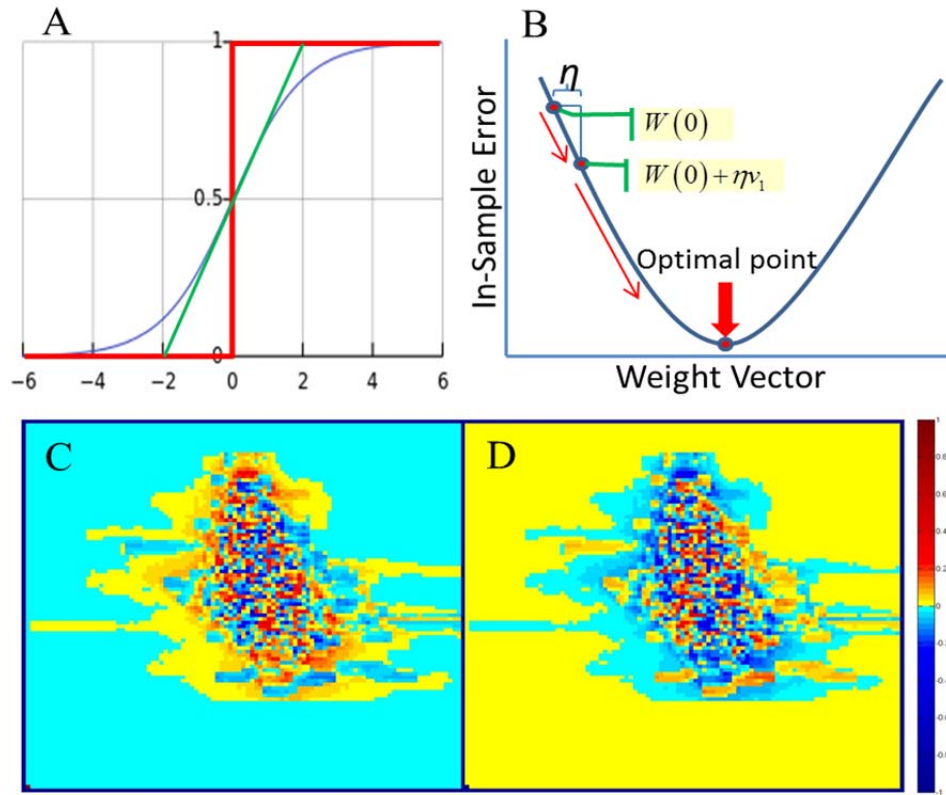
Image A demonstrates different types of thresholds. The red line is a hard threshold giving the results of either 0 or 1. The green line is a linear threshold giving the linear results between 0 and 1. The blue line is a sigmoid threshold giving also the results between 0 and 1 but smoother than a linear threshold. Image B demonstrates the idea of gradient decent. The in-sample error of the weight vector decreases as iteration. It can be viewed as the red ball falls toward the optimal point, which gives minimum error to the weight vector. Images C and D demonstrate the weight matrices (convert from weight vectors) of D-call and 40-Hz call.

Figure 11. Analogies of linear model thresholds, gradient descent, and images of weight matrices for the foraging calls. Adapted from Yaser et al. (2012).

To summarize, the logistic regression algorithm has a default prediction function *h*, which requires a customized weight vector *W* for each targeting object. It initializes *W* at t = 0 as *W(0)*. For t = 1, 2, 3…, n; it computes the gradient g_t and sets the direction $v_t = -g_t$, then updates $W(t+1) = W(t) + \eta v_t$. It iterates to the next step until the designed time to stop then returns the final *W,* which is able to interpret the characteristics of the designated object. Substituting the final *W* for the prediction function *h* allows the function to estimate the probability of input data as each targeted object.

The logistic regression produces two values for an input vector, which, in this research are the probability the image is a representation of a D-call and the probability it is a representation of 40-Hz call. The classification algorithm chooses the highest one to label this input. By comparing the predictions to the annotations, this algorithm provides the classification accuracy for performance estimation.

2.   Requirements for Cross-Validation and Pattern Recognition

The machine learning technique requires adequate data for its learning algorithms. The rule of thumb for any model is "garbage in garbage out," which implies that poor data generate poor results. Most machine learning algorithms have high in-sample accuracy even when the input data are very noisy. That is because these algorithms also learn the characteristics of noise. Knowledge learned from the noise helps these algorithms accurately classify the same data, which is known as "in-sample accuracy." However, ambient noise in different data presents different features. Applying these algorithms, which are trained by a noisy dataset, to another dataset can reveal its true performance, which is the "out-of-sample accuracy." It is difficult, if not impossible, to have a high out-of-sample accuracy when most of the input data are diverse types of noise but not signals. Thus, this research applies pattern recognition to process the acoustic data before using logistic regression, and applies the cross-validation method to estimate the real performance.

a. *Importance of Pattern Recognition*

One hypothesis of this research is that the logistic regression classifier should be able to more accurately classify calls if the inputs are images containing the signal of interest with only minimal ambient noise. Previous researches (Bahoura 2009; Madhusudhana et al. 2010; Bahoura et al. 2012; Helble et al. 2012; Shamir et al. 2014; Karnowski et al. 2015) indicate that performance of any detector or classifier is dominated by ambient noise. The importance of developing an algorithm to filter ambient noise before classifying the foraging calls cannot be overemphasized. Problems of ambient noise are minimized by pattern recognition in this research. The detection algorithm can generate a uniform and very low noise dataset; therefore, it enhances the performance of the classification algorithm.

In addition to filtering ambient noise, pattern recognition may also be applied to detect potential foraging calls from unprocessed acoustic data, which is the hypothesis of developing the selection algorithm. Developing a detector for the foraging calls is even more important since a classifier is useless if there is nothing to classify. The section of the selection algorithm explains the scheme of using pattern recognition to detect foraging calls, and Chapter III provides evidence to support both the hypotheses.

b. *Importance of Cross-Validation*

The logistic regression classifier can provide a very good in-sample accuracy, which is higher than 95%, for classifying the foraging calls within the training subset before applying the pattern recognition technique. The out-of-sample accuracy estimated by our testing subset, however, is less than 50%. The definition of "in-sample" is the samples used to train the classifier. Since this classifier learns everything about these samples, it can accurately classify different objects of these samples. On the other hand, "out-of-sample" means samples that the classifier has never seen before. The classifier applies a prediction function developed from the training subset to classify the testing subset. High in-sample but low out-of-sample performance indicates that this classifier only fits the training subset but cannot generalize characteristics of the foraging calls. This initial experiment emphasizes the importance of the cross-validation method and the

requirement of using multiple subsets to train and to test the classifier. Since out-of-sample accuracy implies the ability of a classifier to correctly classify a new dataset, this research uses the out-of-sample accuracy as an estimate of the performance of the classification algorithm. The number of DCLDE annotated foraging calls is sufficient enough to be partitioned into three subsets. If the initial performance estimated by the testing subset does not satisfy the designed goal (i.e., 95% out-of-sample accuracy in this research), there is a second chance to acquire a higher out-of-sample accuracy since the validation subset is still available.

The classification algorithm is trained by the training subset and generates the initial prediction function. This function is then applied to the training subset for estimating the initial in-sample accuracy; therefore, the thresholds of the classification algorithm such as size of step, initial weight function W, and iteration number of time can be modified. Once the initial in-sample accuracy is high enough, the classification algorithm is applied to the testing subset for initial out-of-sample estimation. This initial estimation provides more clues for modifying the classification and the detection algorithms since out-of-sample accuracy shows the weaknesses of both algorithms. By using both subsets, the loop of modifying and estimating the classification and detection algorithms can run as many times as needed to fit both the algorithms. Surprisingly, increasing the number for iterations to achieve higher in-sample accuracy does not guarantee higher out-of-sample accuracy. Compromise between the performance of detection and classification algorithms is discussed in the next chapter.

D.   PATTERN RECOGNITION FOR CANDIDATE SELECTION

The selection algorithm also applies the pattern recognition technique but aims to select candidates from a series of 5-minute spectrograms, which cover the total duration of the original acoustic data. These candidates include any sound sources with features similar to those of the foraging calls. Using five minutes as an individual period for scanning acoustic data is based on two reasons. First, the general duration of noise from a passing ship is approximately 20 minutes based on the LTSAs of DCLDE 2015 acoustic data. Thus, it requires spectrograms with less than 10 minutes duration (half of the

shipping noise duration) to resolve the shipping noise feature. Another reason is for the future application of this algorithm. These data were collected using a duty cycle that recorded for five minutes, then was off for one or more 5-minute periods in order to extend the deployment time for a fixed number of battery packs.

Determining SNR threshold, modifying call-oriented rules, and creating noise-oriented rules are three concepts used in development of the selection algorithm from the detection algorithm. Both algorithms share four similar major rules. Most of the thresholds for the selection algorithm are adjusted following the three concepts to meet its requirement. The pattern of a foraging call in a 9-second spectrogram is similar to its pattern in a 5-minute spectrogram. Actually, the call itself is the same no matter what size the spectrogram is, if the resolution is the same.

Though the call itself is the same no matter whether the duration of its spectrogram is nine seconds or five minutes, the thresholds of SNR (i.e., percentage of elements selection) for selection and detection algorithms are different. This means the same foraging call may be represented by a different number of elements for different spectrograms. Furthermore, the goals of the two algorithms are different as well. The task of the whole research is similar to finding all the apple trees and pear trees in a forest and labeling them as an apple or a pear tree. An experienced farmer is able to identify both apple and pear trees with a photo taken from 10 meters away from the tree. This skill is learned by the detection and classification algorithms. However, examining every tree in the forest requires a tremendous amount of time. Minimizing the time is the goal of the selection algorithm. The rule of thumb is to keep as many foraging calls as possible since minimizing the time is only the secondary goal of the whole research, but the priority is to detect and classify the foraging calls. The selection algorithm aims to have as high recall (the ratio of detected ground-truth calls to the total ground-truth calls) as possible while the detection algorithm aims to balance the recall and precision (the ratio of correct detections to the total detections). Consequently, the restriction of the call-oriented rules for the selection algorithm requires adjustment to let more potential signals of interest pass through.

Besides minimizing the computational time, another goal of the selection algorithm is to estimate durations of the signals of interest. These durations are used to plot a 9-second spectrogram for each selected candidate (i.e., potential foraging calls) since the detection algorithm is designed to process spectrograms that are centered on foraging calls. Thus, using sequential 9-second spectrograms to represent the acoustic data is not an option for this research, even though the computational time can be reduced by using multiple or powerful computers. The following subsections summarize the four major rules of the selection algorithm with the same sequence of the detection algorithm.

1.   De-Noising and Elements Selection

The task of this rule is to mitigate the effects of ambient noise, which significantly affects the results of the selecting algorithm, and defines the threshold of SNR for a 5-minute spectrogram. Similar to the first rule of the detection algorithm, this rule filters most mechanical noise and selects a limited number of elements between 25 and 90 Hz. The differences are the approaches of filtering ambient noise and the threshold of elements selection.

a.   De-Noising

Noise filtering is always an issue when selecting elements from the spectrogram. This rule is able to filter much of the mechanical noise out and keep as many foraging calls as possible. Most of the DCLDE acoustic datasets have mechanical noises, examples of which are shown in Figure 12. The definition of mechanical noise, which includes tonal and broadband noises, has already been introduced in the detection algorithm subsection. The tonal noise is easy to filter since its frequency band is fixed. Furthermore, filtering this noise only slightly affects the selection results since its bandwidth is less than 5 Hz and the bridging rule is able to fill this gap. Take the CINMS18B_d06_120622_055731 dataset as an example. The mechanic tonal noise is filtered before elements selection by simply filtering all elements between 30 and 32 Hz and 50 and 52 Hz. The frequency bands of mechanic tonal noises in the other datasets may be different, and thus this rule requires adjustment when applying it to different datasets.

The broadband noise, on the other hand, is not easy to filter because of the variability of its strength and time interval. The selection algorithm modifies the noise threshold from the first rule of the detection algorithm to filter the broadband noise. Though the number of temporal elements for a 9-second or a 5-minute spectrogram is different, the number of frequency elements is the same. This rule calculates the number of elements in each temporal column after applying the noise threshold, which is shown in Equation 1. For the detection algorithm, if a column keeps more than 95% of the elements after applying the noise threshold, the column is removed. The selection algorithm increases the noise threshold but removes the column if it keeps 90% of the elements. Although the de-noise function is used in the first rule of the detection algorithm, it is not included in this rule because it aims to filter mechanical noise only.
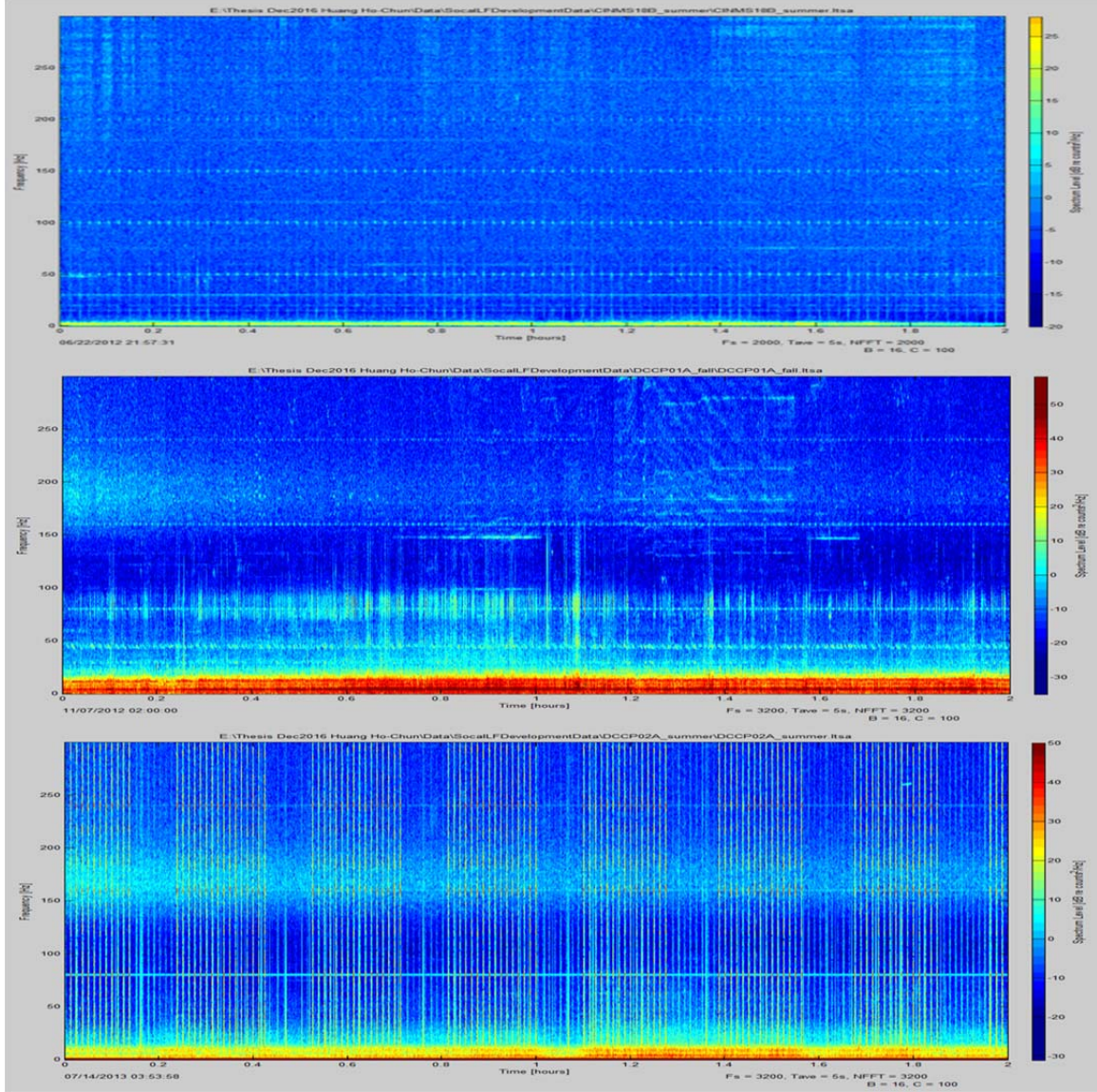
Three 2-hour LTSAs plotted from different locations and seasons have different mechanical noise features. All frequency bands are 0–300 Hz. The top LTSA is plotted from CINM18B_summer acoustic data. The middle LTSA is plotted from DCCP01A_fall acoustic data. The bottom LTSA is plotted from DCCP02A_summer acoustic data. The variety of mechanical noise makes it difficult to apply a uniform rule to all datasets.

Figure 12. Mechanical noise in different datasets.

b.   *Elements Selection*

The SNR threshold determines how many elements are used to represent the candidates. There is indeed a foraging call present in a 9-second spectrogram, which is plotted from annotation and centered on a ground-truth call. However, the continuous

series of 5-minute spectrograms cover all durations of the original acoustic data including periods containing no call or periods containing dozens of calls. A fixed value is no longer adequate and the 4% threshold of the detection algorithm cannot be used. It is better to use a dynamic threshold, which adjusts to the ambient noise and other blue and fin whale vocalizations.

Since the algorithm only selects a limited number of the highest intensity elements, the ambient noes can mask the foraging calls. Furthermore, loud ambient noise may force blue and fin whales to become quiet as suggested by Joseph and Margolina (2015). Thus, elements selected from a noisy period may represent only noise with no foraging calls. The selection algorithm calculates the average intensity between 25 and 90 Hz in a 5-minute period, and then classifies the average intensity values from level 1 to 20 as quiet to loud, as shown in Figure 13. The SNR threshold for level 1 to 15 is 3% and decreases 0.5% per level from level 16 to 20. The threshold only decreases after level 15 because the hypothesis of this algorithm is that the behavior of whales is more likely to change in a noisy environment depending on noise level. Though increasing this threshold also increases the recall, if too many elements are allowed to pass, many noise features would pass and the size of each candidate is also increased. Consequently, each candidate is more likely to connect noise features and the potential foraging call may no longer be in the center of the candidate spectrogram, which makes the detection and classification algorithms cannot recognize the call. Nevertheless, the results show that the recall only increases from 82.33% to 84.26% while the threshold increases from 3% to 6%. This indicates that simply adjusting the threshold by the strength of ambient noise does not provide a satisfactory outcome.

The presence of other blue and fin whales' vocalizations indicates that these whales are currently in the area. Consequently, this period has a high probability for the presence of foraging calls. However, detection of other vocalizations requires more algorithms, which are not included in this research. This discussed in the final chapter as a recommendation for future research.
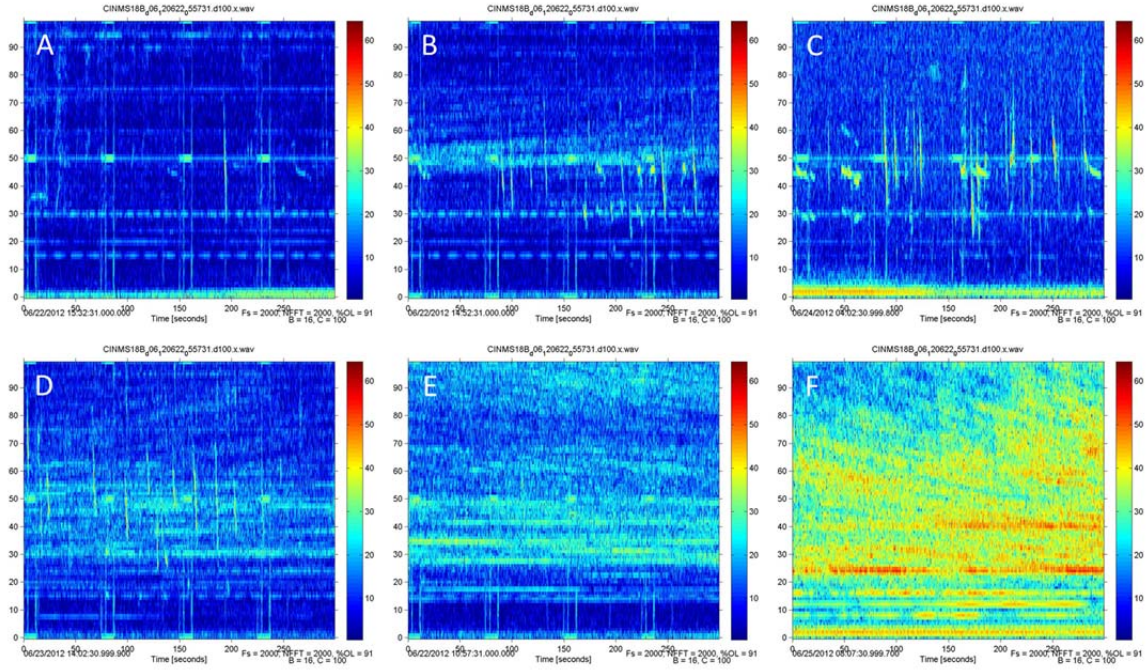
The CINMS18B_d06_120622_055731.d100.x acoustic data are covered by 1342 5-minute spectrograms. Images A, B, C, D, E, and F represent ambient noise level 1, 4, 8, 12, 16, and 20, respectively. These examples imply that the ambient noise does not affect the elements selecting process until it becomes really strong.

Figure 13. Ambient noise level.

This step is critical since the rest of the algorithms become useless once a foraging call is filtered in the beginning. However, evaluating the results of each change requires not only the scoring tool, but also analysts' participation because the DCLDE annotation file does not cover the entire period of DCLDE acoustic data. Furthermore, some ambiguous sound sources can only be evaluated by analysts, but not the scoring tool. An ambiguous sound source is defined as a potential foraging call, which is not annotated, but upon review, is confirmed by a trained analyst as a highly possible foraging call that was overlooked during the annotation process. The examples of ambiguous sound sources are shown in Figure 14. These sound sources make it complicated to estimate the real performance of the selection algorithm, which is discussed in the next chapter.

37
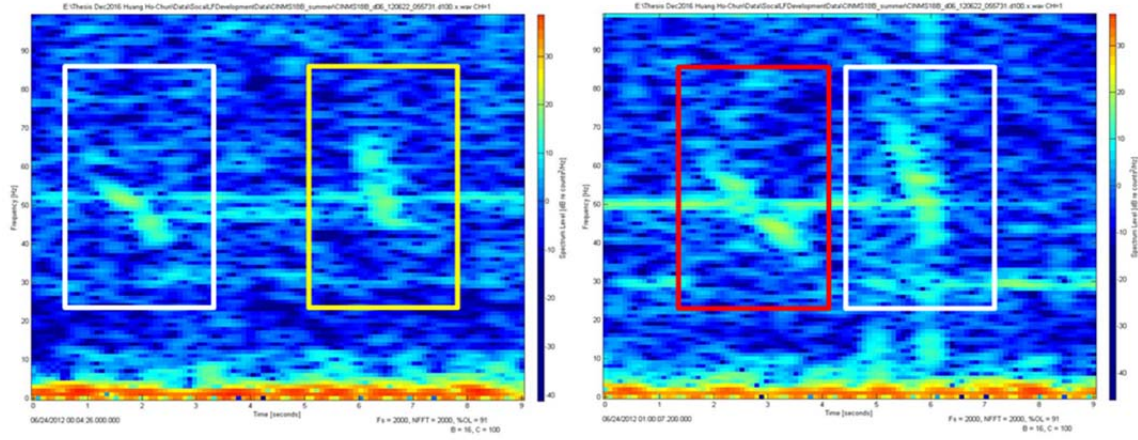
Both images are 9-second spectrograms. The left spectrogram starts from 6/24/2012 00:04:26 and the right spectrogram starts from 6/24/2012 01:00:07. The yellow box indicates an annotated fin whale 40-Hz call. The red box indicates an annotated blue whale D-call. The white boxes indicate two ambiguous sound sources.

Figure 14. Examples of ambiguous sound sources.

2.   Grouping and Dilating

The changes of this rule between the selection algorithm and the detection algorithm are the grouping thresholds. The first grouping threshold for the detection algorithm is six elements but the selection algorithm uses only four elements. Decreasing the threshold makes more small pieces pass through this rule. The second grouping threshold is still 45 elements since this is after dilation. There is an additional grouping threshold in the selection algorithm to filter any group with more than 400 elements. This threshold aims to filter the blue whale A-call and B-call. While the call-oriented rules of both the detection and selection algorithms are similar, the noise-oriented rules are not. Different durations of the spectrogram have different noise patterns (e.g., a blue whale B-call in a 9-second spectrogram is similar to a tonal noise but it is only a small spot in a 5-minute spectrogram), which is shown in Figure 15. The A-call and B-call cannot be filtered by the first rule since they are not like the tonal noise anymore. Though this additional threshold is able to filter the A-call and B-call, it also filters the nearby foraging calls if they connect to each other. The problem is that even applying the rule keeps these big groups because they may contain foraging calls. These groups are too big to fit in a 9-second spectrogram. Furthermore, the centers of the groups are most likely not the centers

38

of the foraging calls. It requires more sophisticated rules, which are not included in this thesis, to separate a foraging call and other sound sources if they connect to each other.

Another issue caused by the A-call and B-call is that their intensity is normally higher than the foraging calls. The elements selected from the 5-minute spectrogram, which contains multiple vocalizations of blue and fin whales, are more likely representing the A-call and B-call but not the foraging calls. The first rule already introduces this issue since the element selection threshold is better to interact with the presence of other blue and fin whale vocalizations.
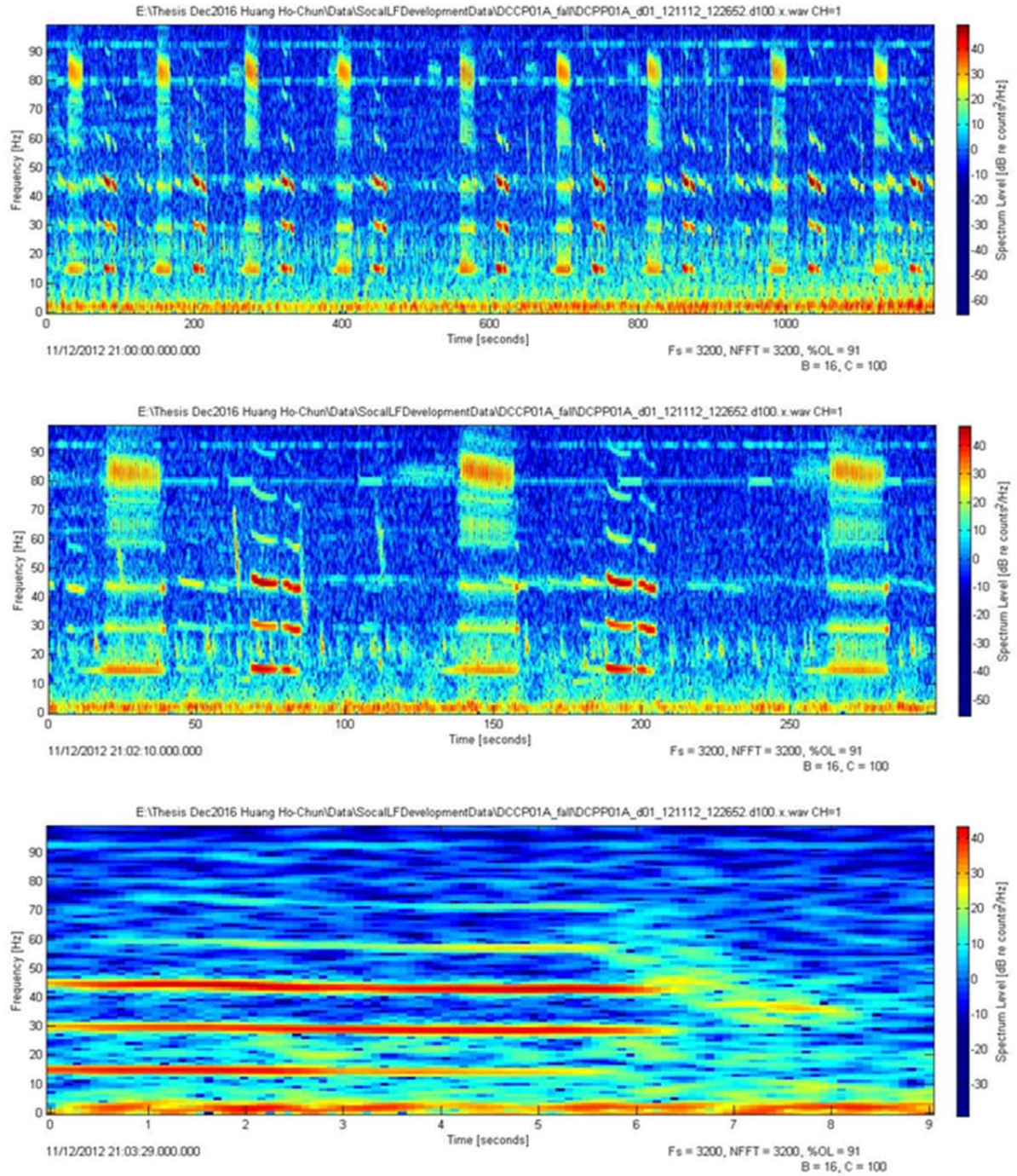
The top image is a 20-minute spectrogram starting from 11/12/2012 21:00:00. The middle image is a 5-minute spectrogram starting from 11/12/2012 21:02:10. The bottom image is a 9-second spectrogram starting from 11/12/2012 21:03:29. The red lines in the 9-second spectrogram belong to a blue whale B-call, which is one of the red spots in other spectrograms. There is a D-call following the B-call. Since the intensity of the B-call is higher than the D-call, selecting elements from the D-call is very difficult.

Figure 15. Noise pattern in different spectrograms.

3. Bridging

The only change of this rule between the selection algorithm and the detection algorithm is the length of the bridge decreasing from 5 to 3 elements. This rule aims to fill the gap from removing the tonal noise with minimum elements. The elements selected from a 5-minute spectrogram sometimes represent many instances of noise but only a few foraging calls. Using a narrower bridge in the selection algorithm can prevent unnecessary connections (i.e., connections of noise to noise or connections of noise to foraging calls). These connections create false alarms and dislocate the foraging call, and thus reduce precision and recall. The results also show that the recall increases as bridge length is decreasing; thus, using 5, 3, and 1 element will provide 0.8426, 0.8480, and 0.8522 recall, respectively.

4. Ridging and Final Assessment

The ridging procedure of this rule is similar to the detection algorithm, but the sets of criteria for determining the candidates are different. The modified first set includes bandwidth $\geq$ 8 Hz, duration $\geq$ 0.45 seconds, ratio of bandwidth to duration $\geq$ 0.33, maximum frequency $\geq$ 35 Hz, and minimum frequency $\leq$ 85 Hz. The modified second set includes duration $\geq$ 1.8 seconds, ratio of bandwidth to duration $\geq$ 0.28, maximum frequency $\geq$ 35 Hz, and minimum frequency $\leq$ 85 Hz. The minimum duration is still 0.45 seconds in the first set because the dilation function increases at least 2 bins for each contour. Thus, a threshold less than 4 bins is meaningless. The upper frequency threshold in both sets increases from 75 to 85 Hz because the selection algorithm is able to filter blue whale A-calls and B-calls, which is the main reason for using 75 Hz in the detection algorithm. The lower frequency threshold does not change because the selection algorithm still cannot filter fin whale 20-Hz calls.

This rule does not have the re-denoise function, which removes the noise tail from a contour for the detection algorithm. It is easy to describe the pattern of a noise tail in a 9-second spectrogram that centered on an annotated foraging call since the call is always in the center, and the noise is only a small portion attached to the left or the right of the call. In a 5-minute spectrogram, however, a similar noise may connect multiple calls and

the pattern of the noise cannot be described by the re-denoise function. To solve this issue requires a new noise-oriented rule, which is not included in this thesis.

Output of the selection algorithm is a file with the same format as the DCLDE annotation file. This output file does not have information for species and call types since the candidates are only potential foraging calls without classification. This file provides temporal information to plot 9-second spectrograms for these candidates, which become the input of the detection algorithm. A uniform format of this file allows the DCLDE scoring tool to estimate the performance of this algorithm, which is discussed in the next chapter.

III. RESULTS AND DISCUSSIONS

The DCLDE 2015 scoring tool is used to quantify the performance of the detection and selection algorithms. The classification algorithm provides its own evaluation for both in-sample and out-of-sample classification accuracy. This chapter introduces the scoring tool and illustrates the individual result of each algorithm.

A.   DCLDE SCORING TOOL

Designed by the Scripps Institution of Oceanography, the DCLDE scoring tool estimates the performance of any detector or classifier. The scoring tool compares the annotation file to the output of the selection and detection algorithms, which are the call candidates and extracted contours, respectively. Their performance is represented by the following scores (the naming convention from the DCLDE scoring tool for the performance metrics has been adopted): precision, recall, percentage of each ground-truth signal that corresponds to one or more detections (truthCoveragePct), average truthCoveragePct for all ground-truth calls (truthCoverageOverallPct), similar to truthCoveragePct for individual detection (detectionCoveragePct), similar to truthCoverageOverallPct for all detections (detectionCoverageOverallPct), and empirical expected fragmentation based on the mean of all fragmentation values that do not correspond to missed ground-truth calls (Efragmentation) (SIO 2015b). The higher the scores, the better—except in the case of Efragmentation. The best Efragmentation value is one, which signifies that each ground-truth call is represented by only one detection but not several broken pieces.

The reason for using precision and recall to represent the performance of a detector or classifier can be illustrated by a confusion matrix, as shown in Figure 16. Since the annotation file contains only foraging calls and the output of the selection and detection algorithms has only positive detections, there is no "true negatives" when comparing annotation to the output of the algorithms. Thus, the DCLDE scoring tool calculates recall and precision but not accuracy.

Although the detection and selection algorithms are able to provide both temporal and frequency information, the scoring tool calculations are only based on temporal information because the DCLDE annotation contains only start time and end time for each ground-truth call without frequency information. The effects of frequency uncertainty are discussed in the following subsection.

$$recall = \frac{TP}{P}$$

$$precision = \frac{TP}{TP + FP}$$

$$accuracy = \frac{TP + TN}{P + N}$$

P is the number of annotated foraging calls. N is the number of "unannotated foraging calls," which does not exist in the DCLDE annotation file. FP can be viewed as the number of detections that do not match the annotation. FN can be viewed as the number of undetected annotations. Since TN does not exist in this case, the DCLDE scoring tool calculates recall and precision but not accuracy.

Figure 16. Confusion matrix and equations of common performance metrics.
Adapted from Fawcett (2006).

B. PERFORMANCE OF DETECTION ALGORITHM

A perfect detector produces no false positive or missed calls and extracts all call contours without noise. However, the reality is that the detection algorithm has to balance all the scores because there is always a tradeoff between each score. Although countless settings were tried while developing the detection algorithm, it is inefficient to compare the effects of all thresholds to demonstrate the procedure for determining the optimal combination. Thus, this subsection introduces the four most sensitive thresholds to illustrate the procedures. The four thresholds are the percentage of element selection, the

targeted frequency band, number of dilations, and bridge length. The other thresholds, which are also important for overall performance, are fixed while comparing these four sensitive thresholds. This process applies the detection algorithm to both training and testing subsets. The final combination of thresholds is used on the validation subset for out-of-sample estimation.

The tradeoff between recall and precision is shown in Figure 17. If the goal is to weigh both values equally, the best combination is 3.2% elements, a frequency band between 25 and 90 Hz, one dilation, and five elements for bridge length. This combination provides the highest F-measure as 0.952 with 0.955 recall, 0.949 precision, 77.7% truthCoverageOverallPct, and 89.7% detectionCoverageOverallPct.

The tradeoff between truthCoverageOverallPct and detectionCoverageOverallPct is shown in Figure 18. If the goal is to weigh both values equally, the best combination is 4% elements, a frequency band between 25and 90 Hz, one dilation, and five elements for bridge length. This combination provides 0.944 F-measure, 0.973 recall, 0.917 precision, 83.5% truthCoverageOverallPct, and 85.0% detectionCoverageOverallPct. The only difference between two combinations is the percentage of elements selected, which shows that the other three thresholds are optimal values. The following paragraphs provide further analysis to determine the threshold of element selection.
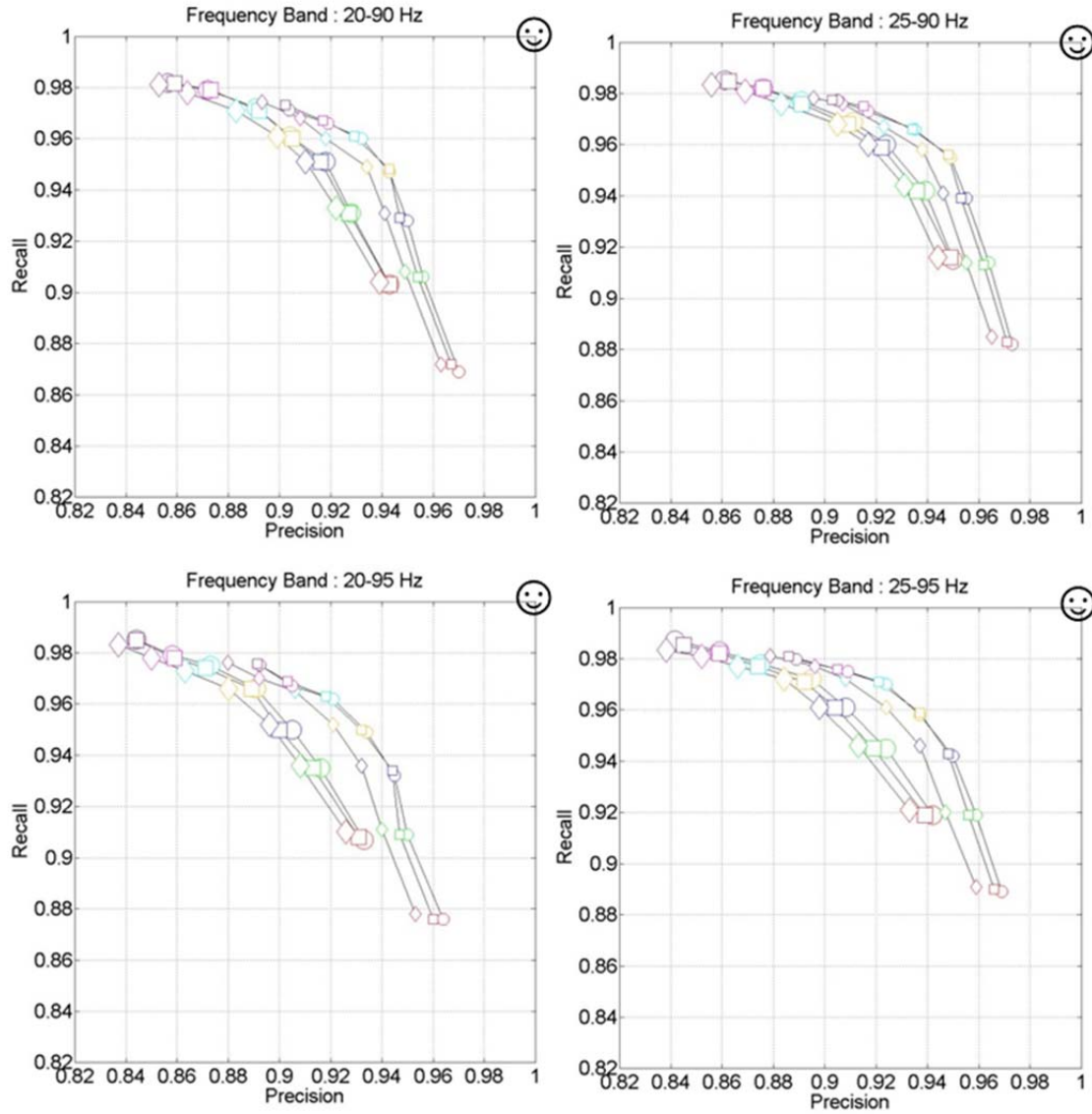
Based on four frequency bands, the symbol's size represents dilation times: small for one time and large for two times. The symbol's shape represents bridge length: circle for five elements, square for seven elements, and diamond for nine elements. The symbol's color represents the percentage of element selection: red, green, blue, yellow, light blue, pink, and purple for 2.0, 2.4, 2.8, 3.2, 3.6, 4.0, and 4.4, respectively. The value of recall increases while precision decreases, and vice versa. The symbol ☺ indicates the ideal point (i.e., perfect performance).

Figure 17. Recall vs. precision.
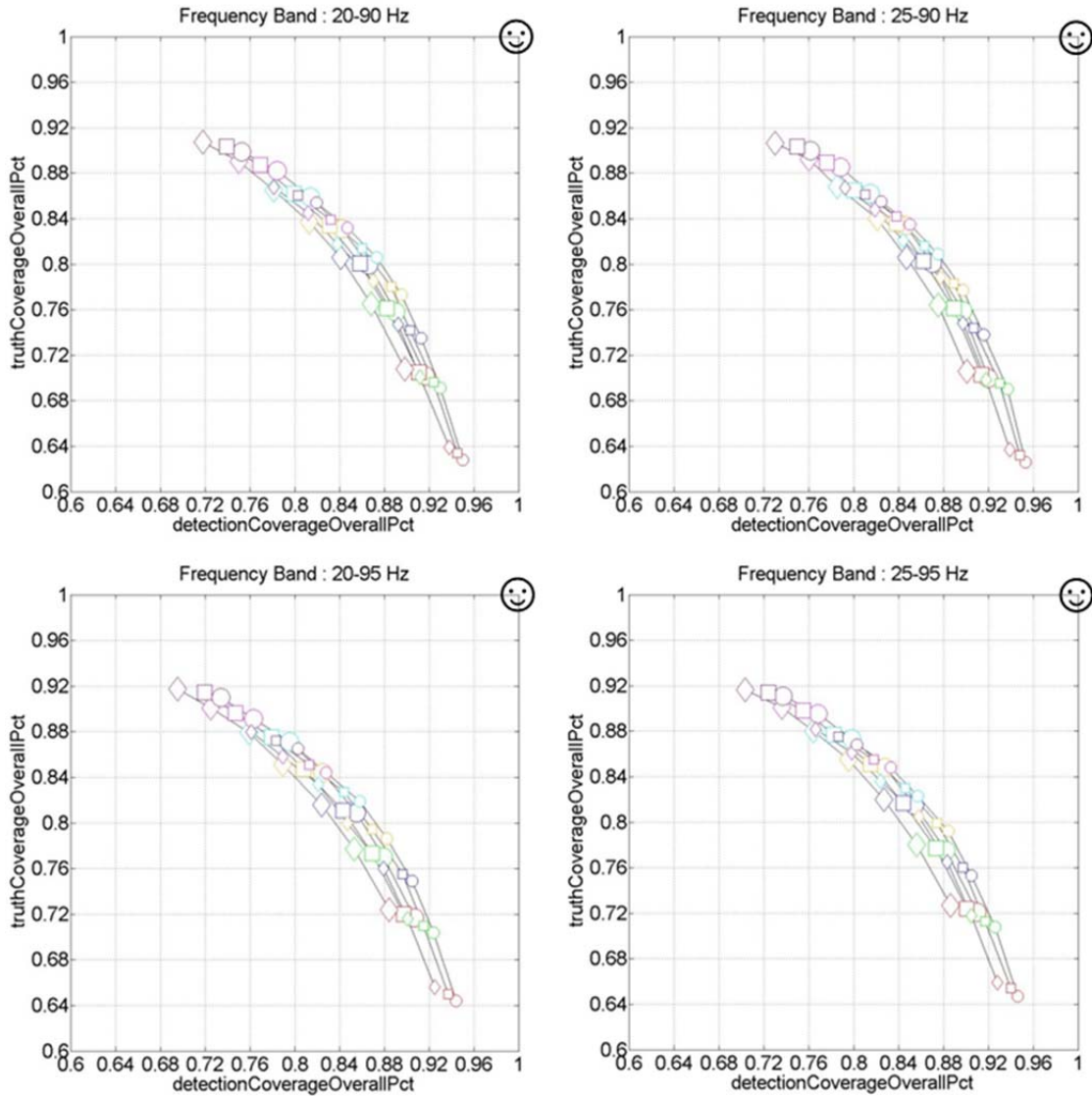
The thresholds of each symbol are the same with Figure 17. The value of truthCoverageOverallPct increases while detectionCoverageOverallPct is decreasing, and vice versa. The symbol ☺ indicates the ideal point (i.e., perfect performance).

Figure 18. TruthCoverageOverallPct vs. DetectionCoverageOverallPct.

While using the scoring tool is a standardized method to estimate the detection algorithm, there are two biases in this process that require further clarification. The first bias is due to the cross-validation method. Each 9-second spectrogram is centered on one annotated call and ideally contains only one call. However, two short-interval calls can appear in one spectrogram once they are close enough to each other. If the detection algorithm detects both calls but one of them is assigned (as shown in Figure 19A) to the

validation subset, this one would be identified as a false positive due to the scoring tool's inability to pair the detection information to validation annotation. Additionally, there may be two overlapping spectrograms that both contain two ground-truth calls (as shown in Figures 19B, 19C). The scoring tool, on the other hand, determines the second (or more) detection as a fragmentation if both calls are assigned to the same subset. This bias implies that the real precision value is slightly higher (a few false positives are actually true positives) and the Efragmentation is slightly lower (a few fragmentations are the other calls but not broken parts of a call) than the numbers calculated by the scoring tool.
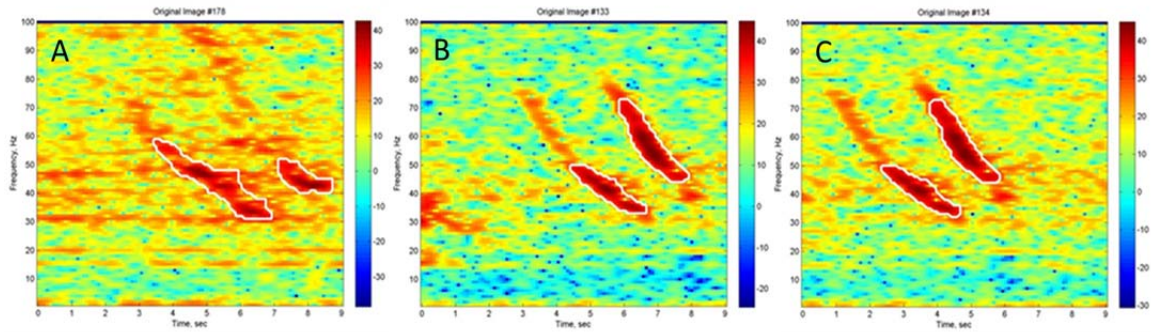

Images A, B, and C are 9-second spectrograms for #178, #133, and #134 annotated calls in the in-sample subset (combining training and testing subsets). Image A shows that two ground truth D-calls are correctly detected (white outline) but the right one is assigned to the validation subset. This situation makes the scoring tool determines the right contour as a false positive. The other two images show that two ground truth D-calls are correctly detected twice because they are too close to each other; therefore, the spectrogram of each call contains both of them. Additionally, their duration overlaps with each other, and the scoring tool determines that each call is represented by four detections and the fragmentation of each call becomes four.

Figure 19. Biases of true positives.

Another bias is due to the lack of frequency information. The scoring tool uses only temporal information to identify detections and, therefore, cannot reliably distinguish between correct and false detections by frequency content. A false detection occurs when the algorithm catches a noise feature that is of the same time duration as the call but occurs at a different frequency band. If the annotated call is also detected (as shown in Figure 20A), both Efragmentation and precision increase. However, if the ground-truth call is missed (as shown in Figure 20B), both precision and recall increase.

48

This bias means that the Efragmentation, precision, and recall values are actually lower than those calculated by the scoring tool.
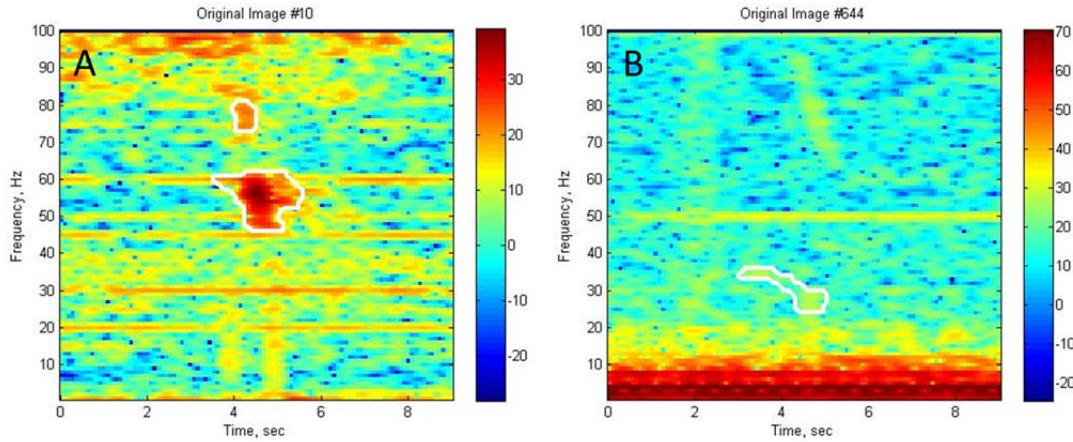
Left: A noise (upper white outline contour) overlaps a ground-truth call (lower white outline contour) in time. This makes the scoring tool determine them as two broken parts of a call. Right: A noise (white outline contour) overlaps a ground-truth call (missed by the detection algorithm but should be the yellow portion in the upper center of the spectrogram). The scoring tool determines it as a correct detection, which accidentally increases the recall.

Figure 20. Bias of false positives.

Although these biases imply that actual recall and precision may be lower than the numbers calculated by the scoring tool, the previously introduced "ambiguous sound sources" implies that precision may actually be higher. Furthermore, better truthCoverageOverallPct and detectionCoverageOverallPct indicate the foraging calls are detected more precisely. Thus, the 4% threshold for elements selection is considered better than a 3.2% threshold. The final combination of four sensitive thresholds is 4% of the elements, a frequency band between 25 and 90 Hz, one dilation, and five elements for bridge length (as shown in Table 1). The final scores of the detection algorithm for its in-sample and out-of-sample performance are shown in Table 2.

49

Table 2.   Final performance of the detection algorithm.

|	In-sample performance (based on 3860 calls in the training and testing subsets)	Out-of-sample performance (based on 964 calls in the validation subset)
Precision	0.92256	0.9203 (0.9166)
Recall	0.9710	0.9668 (0.9627)
TruthCoverageOverallPct	83.35	82.68
DetectionCoverageOverallPct	85.20	84.41
Efragmentation	1.0774	1.0655

Blue color indicates the modified values.

The imperfections of the detection algorithm are revealed when analyzing all contours extracted from the validation subset. There are 32 calls missed since the recall is 0.9668 for the out-of-sample performance. Twenty-four of them are partly extracted but do not pass the final rule. Three examples of partially extracted calls are shown in Figure 21. The first example (Figure 21A) shows a slight upsweep in frequency, which is in contrast to the normal feature of the foraging calls. The second example (Figure 21B) shows the difficulty of extracting a blurry call, which even a well-trained analyst cannot identify with certainty. The third example (Figure 21C) shows the effects of broadband noise when the spectrogram is plotted for a 2.4-second blue whale D-call but most of the call is masked by noise. Even though part of the call is detected, no contour passes the final assessment.
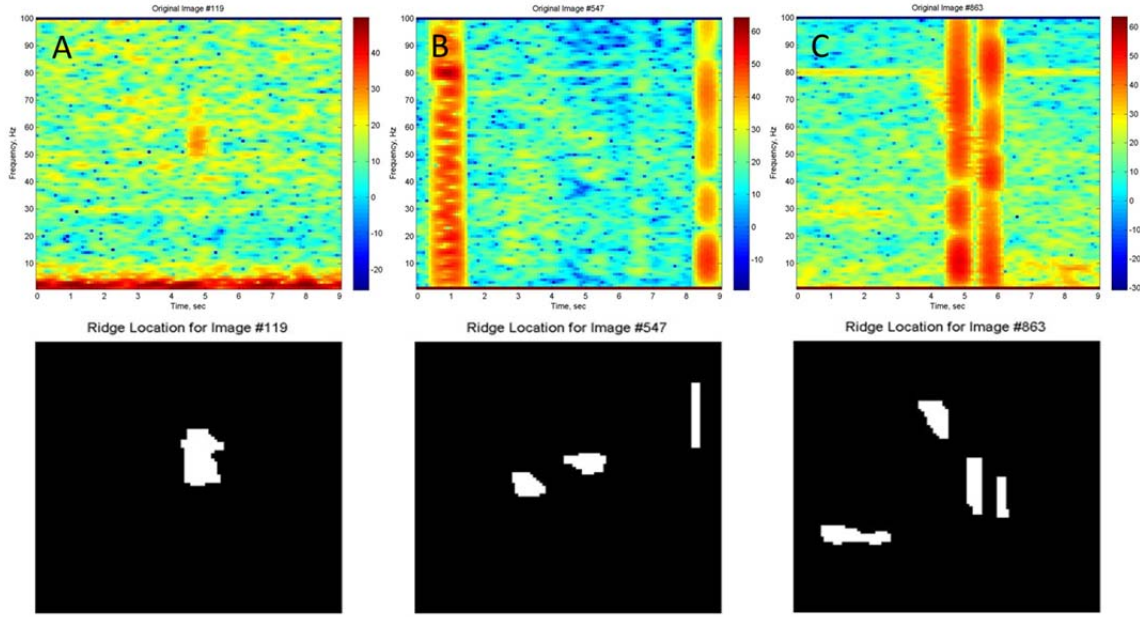
Images A, B, and C are 9-second spectrograms for #119, #547, and #863 annotated calls in the validation subset. The bottom images are their element groups (white contours) before applying the final rule. All contours do not meet the criteria. Thus, the detection algorithm cannot identify them as foraging calls.

Figure 21. Unqualified call contours.

The other eight missed calls are filtered by the first two rules, which can be explained by two reasons. First, a small (short duration and narrow bandwidth) and weak (low intensity) call close to a large and strong call (as shown in Figure 22A) cannot be detected. The 4% elements rule is used to represent the strong call. Thus, the weak one totally disappears. Second, some calls are totally masked by broadband noise (e.g., an annotated 1.5-second blue whale D-call is completely masked by broadband noise as shown in Figure 22B). In addition to these 32 missed calls, four false positives are mistakenly determined as true positives since their durations overlap four undetected annotated calls. After manually analyzing all the out-of-sample results, the modified scores are 0.9166 precision and 0.9627 recall, as shown in Table 2 (blue color).
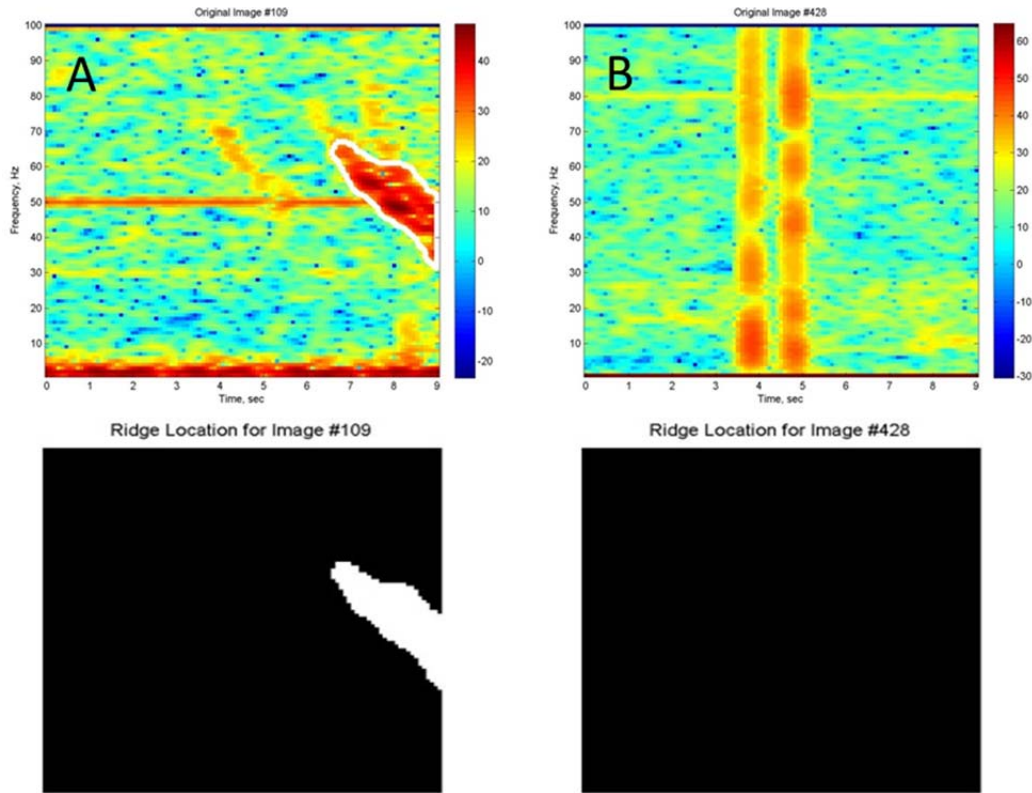
Images A and B are 9-second spectrograms for #109 and #428 annotated calls in the validation subset. The bottom images are their element groups before applying the final rule. The white line in image A indicates that the contour is identified as a foraging call by the detection algorithm but this contour does not represent the annotated call. Elements of both annotated calls cannot pass the first two rules. Thus, the bottom images do not show any relative contour.

Figure 22. Totally missed foraging calls.

While manually reviewing recall and precision helps to better estimate the performance of the detection algorithm, reviewing truthCoverageOverallPct and detectionCoverageOverallPct is too subjective because of the ambiguity of the annotated call durations. This ambiguity can be explained by three factors. First, durations measured from calls' spectrograms or from their time series are different, as shown in Figure 23. Second, using different NFFTs provides different resolutions and shows different start and end times for the same call, as shown in Figure 23. Finally, the subjectivity of an individual analyst cannot be quantified, especially when examining blurry calls (Figure 21B) or masked calls (Figure 21C, Figure 22B, and Figure 24). These three factors suggest that to manually review the truthCoverageOverallPct and

detectionCoverageOverallPct is inadequate and that 82.68% truthCoverageOverallPct and 84.41% detectionCoverageOverallPct are quite precise.
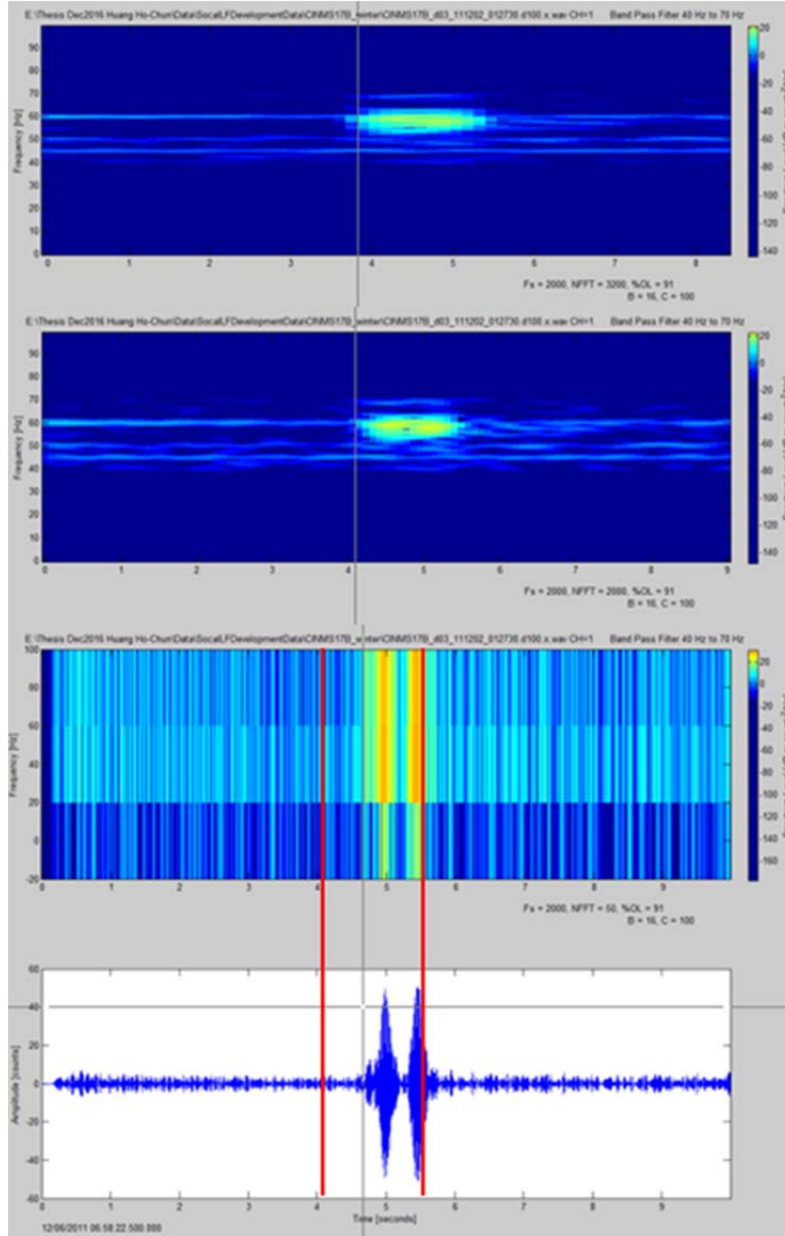
The duration of the annotated fin whale 40-Hz call in this image is form 12/06/2011 06:58:26.6 to 12/06/2011 06:58:28.0 (indicating by red line). All images use a band filter on 40–70 Hz for better analyzation of time series. The FS is 2000 and the NFFT used for each spectrogram from top to bottom is 3200, 2000, and 50, respectively. The start and end time observed from different spectrograms will be different because of different resolutions.

Figure 23. Ambiguity of annotated call durations.

53

C.   PERFORMANCE OF CLASSIFICATION ALGORITHM

The final in-sample and out-of-sample classification accuracies are 98.08% and 95.87%, respectively. The high out-of-sample performance indicates that this algorithm has the ability to classify blue whale D-calls and fin whale 40-Hz calls from a new dataset. Analysis of the input data for the classification algorithm helps further explain its performance.

The classification is trained by the contours extracted from the in-sample dataset, which includes training and testing subsets. The hypothesis of the training process assumes that each contour contains features of either a blue whale D-call or a fin whale 40-Hz call. If only a portion of the call is extracted or the extracted contour is a noise but not the annotated call, this confusion will mislead the classification algorithm and increase the number of misclassifications. This input bias is similar to the bias of the scoring tool, which accidentally treats a noise contour as a ground-truth call because of the lack of frequency information. Another input bias is mainly caused by ambient noise, which covers most features of some ground-truth calls, as shown in Figure 24. The features of these calls cannot be observed from the spectrograms without imagination, which the detection algorithm does not have yet. The detection algorithm can only detect a portion of these calls and labeled them as D-calls according to the annotation. These input biases mislead the classification algorithm in both the training and evaluation processes.

The in-sample performance is based on 3,640 input images. These images include 3,418 D-calls and 222 40-Hz calls. There are 70 in-sample misclassifications including 21 D-calls and 49 40-Hz calls, as shown in Figure 25. The out-of-sample performance is based on 895 input images including 842 D-calls and 53 40-Hz calls. There are 37 out-of-sample misclassifications including 16 D-calls and 21 40-Hz calls, as shown in Figure 26. The individual in-sample and out-of-sample classification accuracies for 40-Hz calls are 77.93% and 60.38%, respectively. These results imply that the classification algorithm needs to learn more from fin whale 40-Hz calls given that the in-sample dataset only provides 222 samples (6% of total samples), as shown in Figure 27. The weight vector (W) for 40-Hz calls in the classification algorithm learns the energy distribution from

these samples. The lack of samples limits the algorithm's learning experience and reduces its ability to correctly calculate the probability of a detection as a 40-Hz call. Therefore, it is believed that the performance of the classification algorithm can be improved by using more training samples of 40-Hz calls.
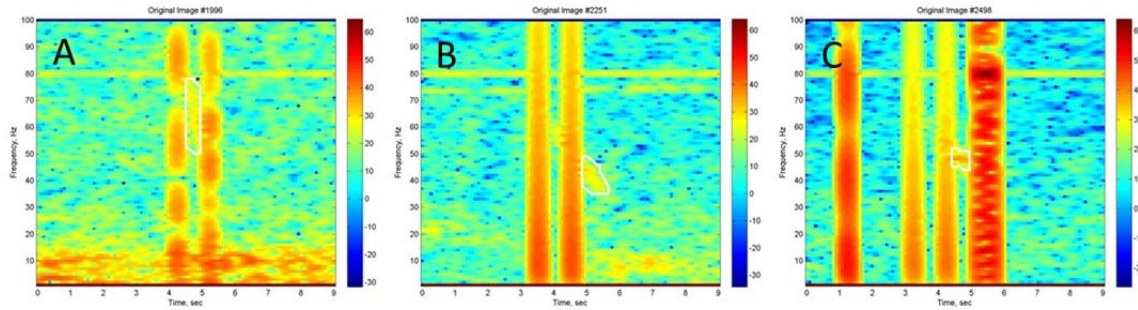
Image A, B, and C are 9-second spectrograms for annotated calls #1996, #2251, and #2498 in the in-sample dataset al.l of them are D-calls and their durations are annotated as 0.8, 1.8, and 2.8 seconds, respectively. The white outline indicates their extracted contours, which are the inputs of the training process for the classification algorithm. These inputs are labeled as D-calls according to the annotation, but have short duration and narrow bandwidth, which are more similar to 40-Hz calls.

Figure 24. D-calls masked by ambient noise.

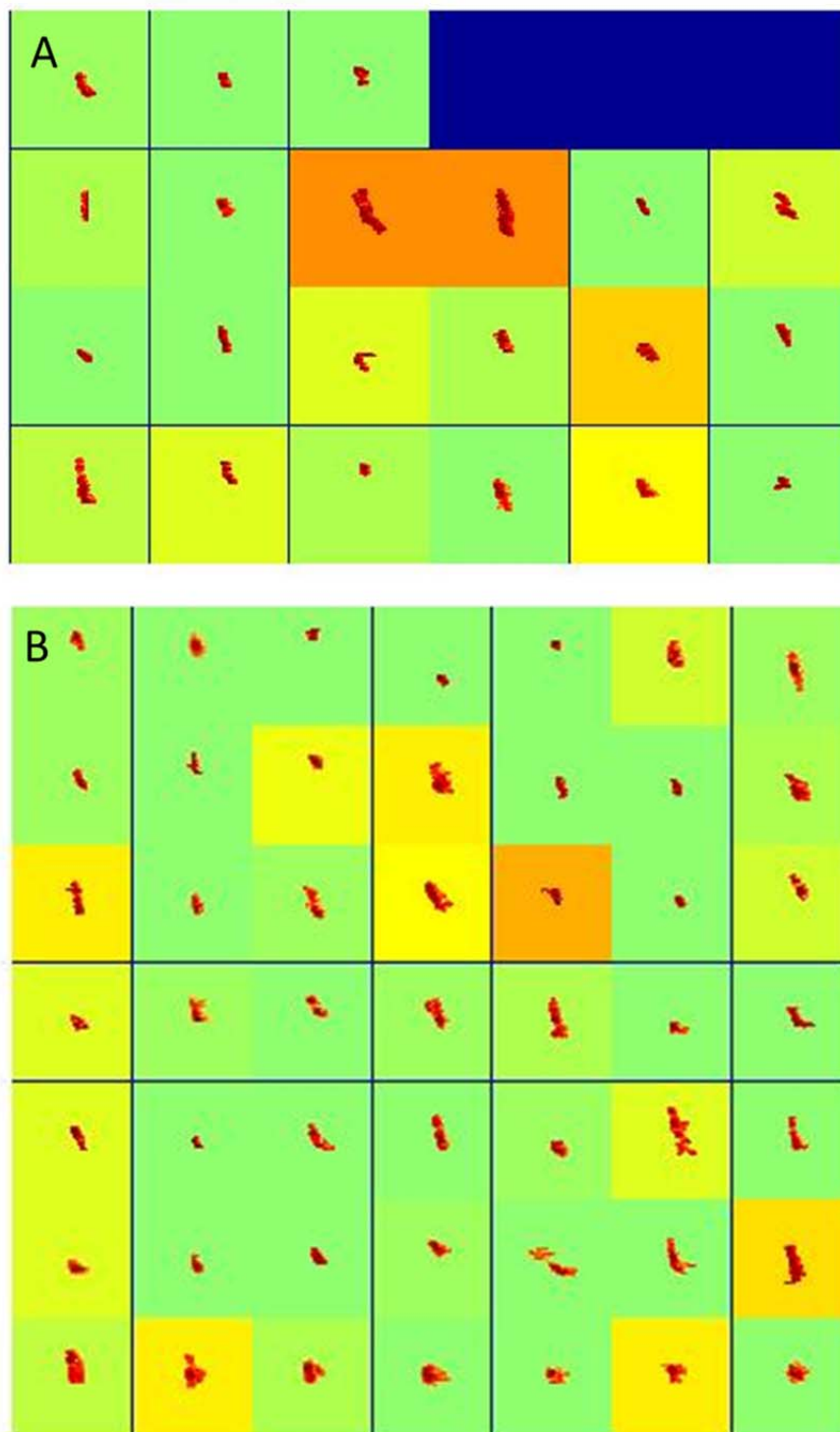Image A includes 21 in-sample misclassified D-calls. Image B includes 49 in-sample misclassified 40-Hz calls.
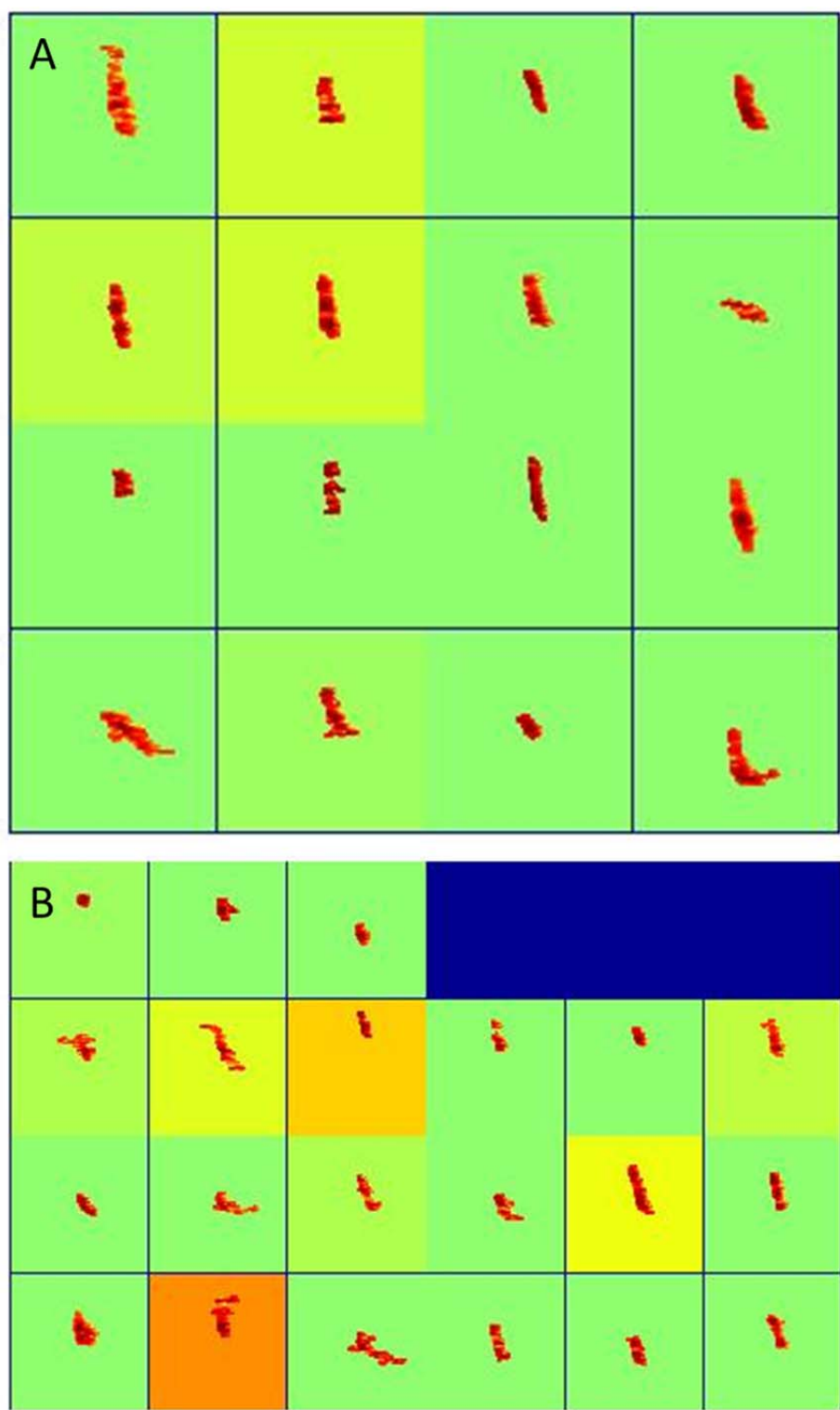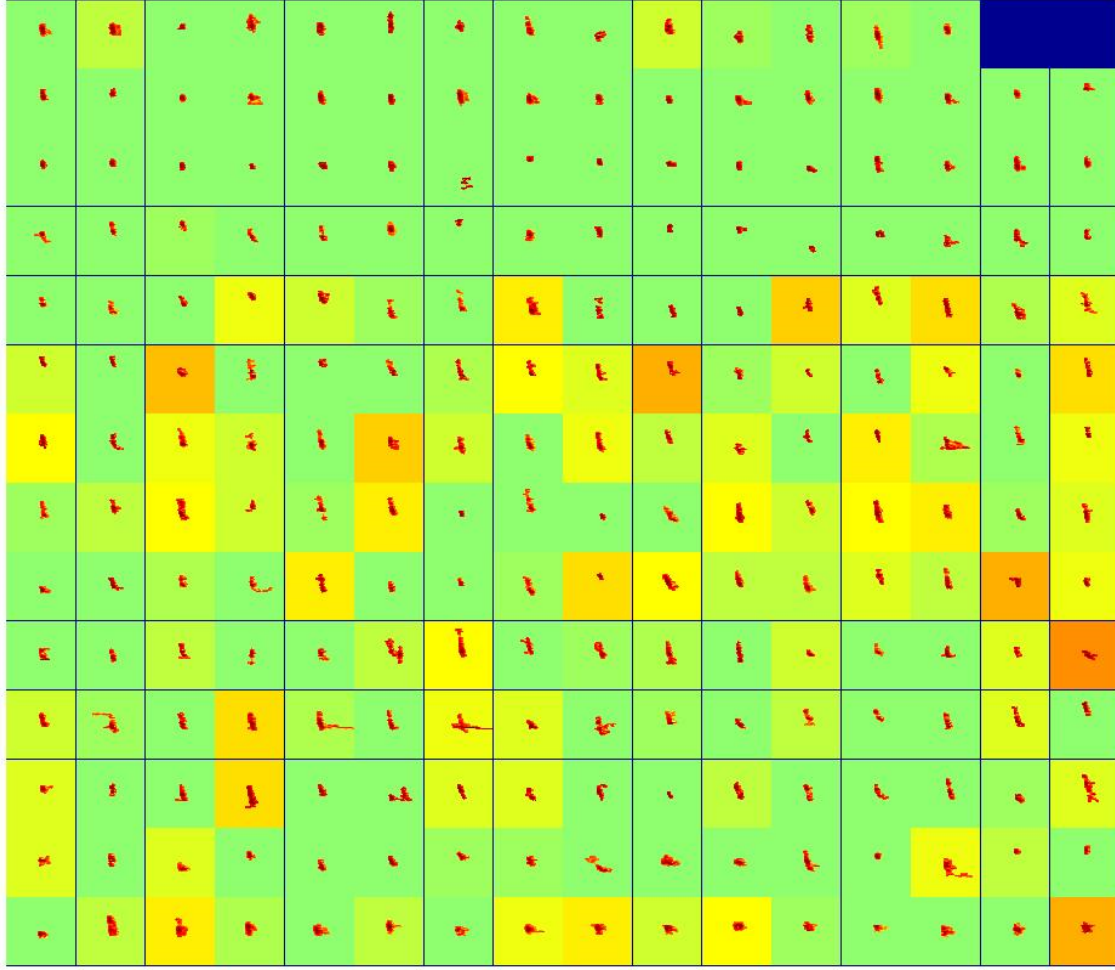
Figure 25. In-sample misclassifications.

Image A includes 16 out-of-sample misclassified D-calls. Image B includes 21 out-of-sample misclassified 40-Hz calls.

Figure 26. Out-of-sample misclassifications.

This image includes 222 fin whale 40-Hz calls in the training samples. There are 3,640 training samples, which means only 6% of them are 40-Hz calls. This image also indicates that some of the samples are masked by ambient noise. The effects of the ambient noise are also enlarged because of lack of training samples.

Figure 27. Training samples for 40-Hz calls.

D.   PERFORMANCE OF SELECTION ALGORITHM

The current performance of the selection algorithm is 0.8522 recall, which is calculated by the scoring tool only for the CINMS18B_d06_120622_055731.d100.x acoustic data. However, the scoring tool cannot properly evaluate the selection algorithm not only because the annotation does not cover all of the acoustic data, but also because of the subjectivity of the annotation. The duration of this acoustic data is 111.83 hours (from 22/6/2012 05:57:31 to 26/6/2012 21:47:31), but the annotation only covers 92

hours. There are 934 annotated foraging calls during these 92 hours, and the selection algorithm chooses 3,145 candidates. One-hundred and sixty-eight of the 934 annotated calls are not selected; however, 185 candidates that are not shown on the annotation are possible foraging calls determined by a post analysis conducted after the annotation was completed. The best explanation for both the undetected calls and the possible foraging calls is the uncertainty of human performance. Thus, some blurry calls are annotated as the foraging calls, but others are not. Examples of annotated blurry calls are shown in Figure 28, and the potential unannotated foraging calls are shown in Figure 29. All examples shown in Figure 28 and Figure 29 are detected from the 92 hours. Since there is no analytical answer for the question of why some are determined as foraging calls but others are not, a new hypothesis is used to estimate the performance of the selection algorithm. This hypothesis assumes that there are 1,119 ground truth calls, including 934 annotated calls and 185 potential unannotated foraging calls. The recalls for the annotation and the selection algorithm are 0.8347 (934/1119) and 0.8499 ((934-168+185)/1119), respectively. This hypothesis implies that although the selection algorithm cannot detect all annotated calls, its performance may be better than that of a human. Furthermore, the same results from the same data can be expected for the algorithm every time, no matter the size of data, the computer used, or when the algorithm is used. However, it is difficult, if not impossible, to obtain the same results from multiple analysts who process weeks or even months of acoustic data.

The selection algorithm pinpoints 1,153 candidates from the uncovered period (22/6/2012 05:57:31 to 23/6/2012 00:00:00), and within 20 minutes, an experienced analyst confirms 457 of them are foraging calls. It is much more efficient for an analyst to identify uniform images, which are 9-second spectrograms plotted for candidates, than to use traditional protocol to annotate foraging calls from unprocessed acoustic data. It is very easy to lose track of calls if there are multiple calls in the LTSA. Furthermore, analysts have to readjust the frequency band each and every time when zooming in on the next signal from LTSA. Additionally, logging the start and end times for a call is also a time-consuming task. None of these chores bothers the analyst when applying the selection algorithm. With help from the selection algorithm, an annotation that contains

457 foraging calls is created within 20 minutes. Although many improvements can be made, the current results indicate that using the pattern recognition technique as the selection algorithm is practical. This algorithm can help analysts efficiently annotate more ground truth samples, which are critical for developing any automated detector and classifier. Eventually, the combination of selection, detection, and classification algorithms will be a reliable detector and classifier will be as well once more noise-oriented rules are developed and more samples of 40-Hz calls are obtained.
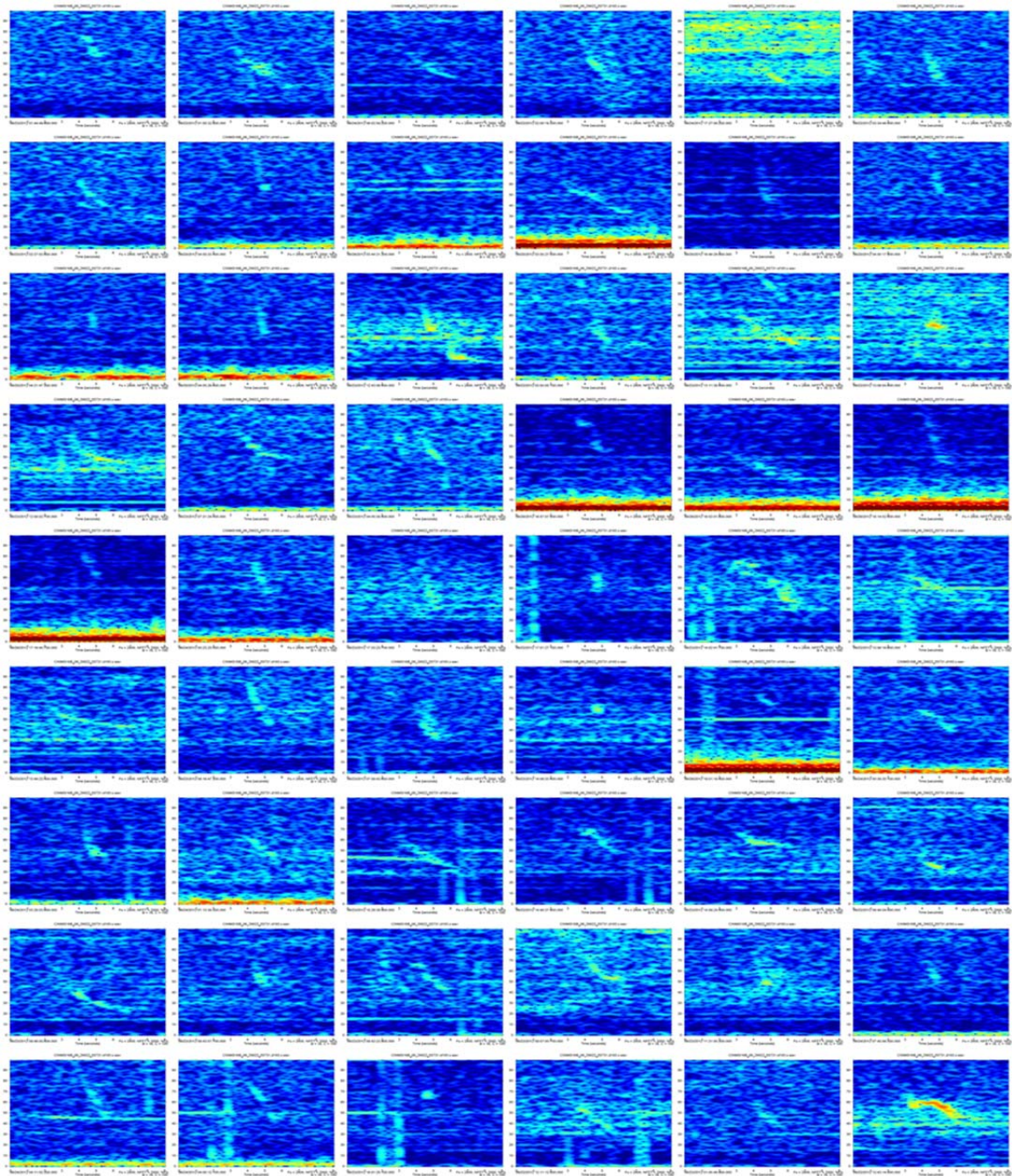
This image includes 54 annotated calls, which are plotted in the center of their 9-second spectrograms, from the CINMS18B_d06_120622_055731.d100.x acoustic data.

Figure 28. Blurry annotated foraging calls.
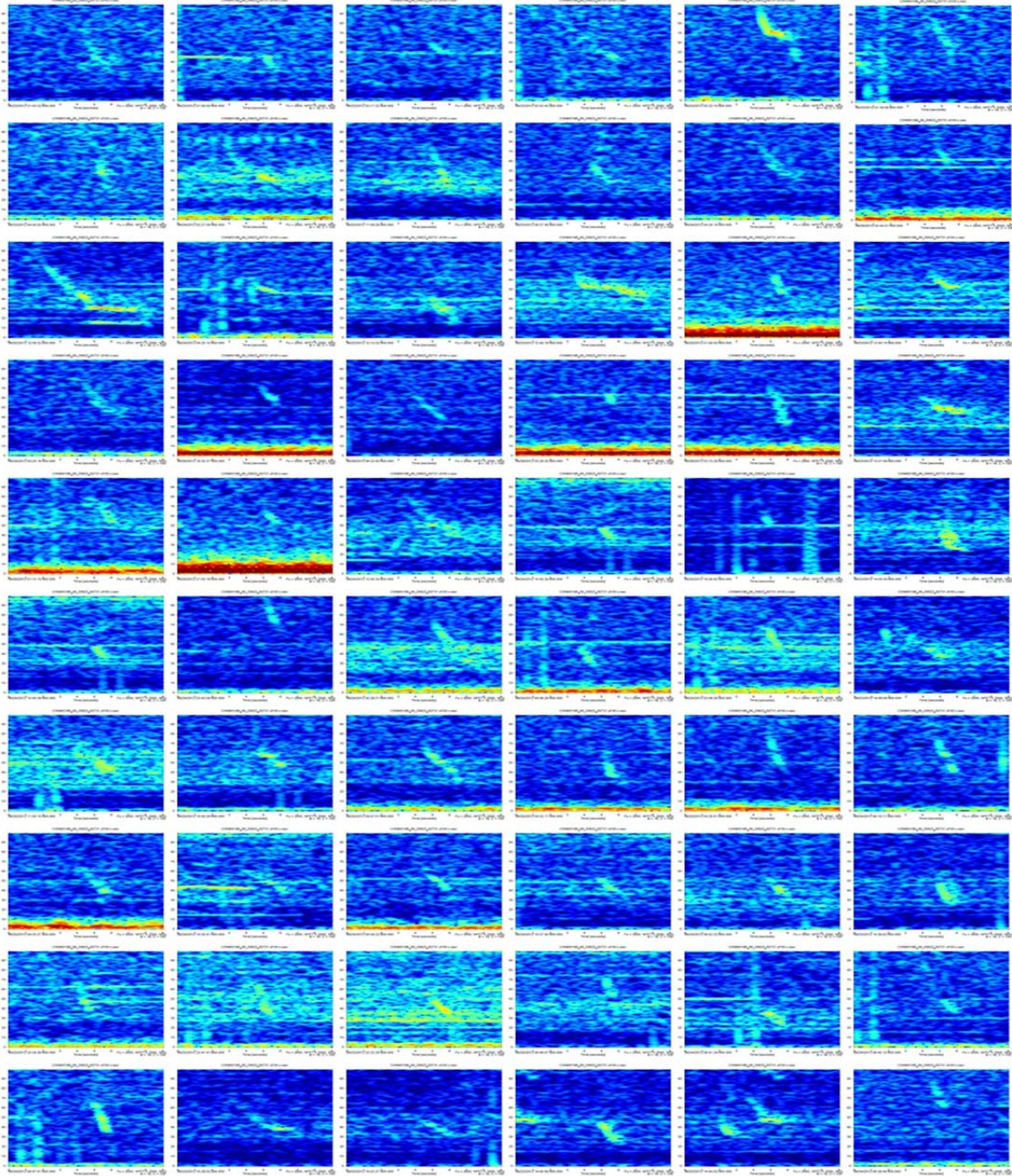
This image includes 60 unannotated signals, which are plotted in the center of their 9-second spectrograms, from the same period for Figure 26. There is no analytical answer to explain why signals of Figure 26 are the foraging calls but signals in this image are not.

Figure 29. Potential unannotated foraging calls.

IV. CONCLUSIONS AND RECOMMENDATIONS

The DCLDE scoring tool and annotation data were used to estimate the performance of the new approach. The detection algorithm, which applies the pattern recognition technique, provided a 0.9627 out-of-sample recall. The classification algorithm, which applies machine learning technique, provided 95.87% out-of-sample classification accuracy. The selection algorithm is still in progress, but also applies the pattern recognition technique and showed potential for detecting more foraging calls from the acoustic data than those annotated by analysts.

The results indicated that more efforts should be dedicated to improve this approach. Although the current protocol of applying three algorithms on raw acoustic data without supervision provides many false alarms, the combination of this approach and manual supervision (as shown in Figure 30) can make the detection and classification of the foraging calls more efficient. With further improvement, this approach has potential to become totally automated and requires minimum supervision to assure quality control.

A. ADVANTAGES AND BENEFITS OF THE AUTOMATED DETECTOR AND CLASSIFIER

While more analysis is required, the key advantages of this automated detector and classifier over the traditional manual approach are its reproducibility, known performance, cost-efficiency, and automation. Even the most well-trained analysts cannot provide uniform performance. Thus, the accuracy of any annotation data that is used as ground-truth performance is difficult to quantify. Moreover, the development of recording techniques provides numerous data, making manual processing of data unpractical. Additionally, there are many applications for using autonomous mobile platforms (e.g., ocean gliders) to detect marine mammals in real-time (Baumgartner et al. 2013). The real-time monitoring requires a combination of platforms, acoustic recorders, low-power digital processors, satellite communications, and efficient detectors like the newly designed algorithms in this research. In sum, this automated detector and classifier

has the potential to provide consistent results efficiently and effectively, making it applicable to huge data processing and real-time monitoring of the baleen whale foraging calls.

B. SUGGESTIONS FOR FUTURE RESEARCH

It is suggested to apply this approach to additional acoustic datasets to further examine the algorithms and efficiently create more ground-truth samples. Meanwhile, the selection algorithm requires more efforts to develop a dynamic threshold for element selecting and more rules for de-noising. The classification algorithm requires more samples of fin whale 40-Hz calls.

Add frequency information in the future, e.g., Bandwidth 45-73 Hz.

Project	Site	Species	Start Time	End Time	Call Type
CINMS	B	Bp	2012-06-23T23:47:27.1	2012-06-23T23:47:28.4	40Hz
CINMS	B	Bp	2012-06-23T23:49:18.1	2012-06-23T23:49:19.3	40Hz
CINMS	B	Bm	2012-06-23T23:49:58.6	2012-06-23T23:50:01.1	D
CINMS	B	Bm	2012-06-23T23:51:32.6	2012-06-23T23:51:35.9	D
CINMS	B	Bp	2012-06-23T23:52:12.3	2012-06-23T23:52:14.3	40Hz
CINMS	B	Bm	2012-06-23T23:53:19.0	2012-06-23T23:53:21.8	D

Balaenoptera musculus (Bm)- blue whale
Balaenoptera physalus (Bp) - fin whale

Current reliability of this automated detector and classifier is not high enough to operate without supervisions. However, the supervisions are eventually unnecessary after more modifications for the algorithms.

Figure 30. Application flow chart of the new approach.

APPENDIX. COMBINATIONS OF DILATION MATRIX

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Bahoura, M., 2009: Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes. *Computers in Biology and Medicine*, **39**, 824–843, doi:10.1016/j.compbiomed.2009.06.011.

Bahoura, M., Y. Simard, 2012: Serial combination of multiple classifiers for automatic blue whale calls recognition. *Expert Systems with Applications*, **39**, 9986–9993, doi:10.1016/j.eswa.2012.01.156.

Baumgartner, M. F., D. M. Fratantoni, T. P. Hurst, M. W. Brown, T. V. N. Cole, S. M. Van Parijs, and M. Johnson, 2013: Real-time reporting of baleen whale passive acoustic detections from ocean gliders. *Journal of the Acoustical Society of America*, **134**, 1814.

Branch, T. A., K. Matsuoka, and T. Miyashita, 2004: Evidence for increases in Antarctic blue whales based on Bayesian modelling. *Marine Mammal Science*, **20**, 726–754, doi:10.1111/j.1748-7692.2004.tb01190.x.

Branch, T. A., and Coauthors, 2007: Past and present distribution, densities and movements of blue whales Balaenoptera musculus in the Southern Hemisphere and northern Indian Ocean. *Mammal Review*, **37**, 116–175, doi:10.1111/j.1365-2907.2007.00106.x.

Castellote, M., C. W. Clark, and M. O. Lammers, 2012: Acoustic and behavioural changes by fin whales (Balaenoptera physalus) in response to shipping and airgun noise. *Biological Conservation*, **147**, 115–122, doi:10.1016/j.biocon.2011.12.021.

Clark, C. W., J. F. Borsani, and G. Notarbartolo-Di-sciara, 2002: Vocal activity of fin whales, Baalaenoptera physalus, in the Ligurian Sea. *Marine Mammal Science*, **18**, 286–295, doi:10.1111/j.1748-7692.2002.tb01035.x.

Croll, D. A., J. Gedamke, A. Acevedo, C. W. Clark, J. Urban, S. Flores, and B. Tershy, 2002: Bioacoustics only male fin whales sing loud songs. *Nature*, **417**, 809, doi:10.1038/417809a.

Edds, P. L., 1988: Characteristics of finback Balaenoptera physalus vocalizations in the St. Lawrence Estuary. *Bioacoustics*, **1**, 131–149, doi:10.1080/09524622.1988.9753087.

Helble, T. A., G. R. Ierley, G. L. D'Spain, M. A. Roch, and J. A. Hildebrand, 2012: A generalized power-law detection algorithm for humpback whale vocalizations. *Journal of the Acoustical Society of America*, **131**, 2682–2699, doi:10.1121/1.3685790.

Huang, H. C., J. E. Joseph, M. J. Huang, and T. Margolina, 2016: Automated Detection and Identification of Blue and Fin Whale Foraging Calls by Combining Pattern Recognition and Machine Learning Techniques. *Proc. IEEE Int. Conf. on OCEANS 2016 MTS/IEEE Monterey, 2016*, Monterey, CA, IEEE and MTS, doi:10.1109/OCEANS.2016.7761269.

Joseph, J. E., and T. Margolina, 2015: Quantifying response in vocal behavior of fin whales to local shipping in the Southern California, *Program Abstracts, 169th Meeting of the Acoustical Society of America*, Pittsburgh, PA, *Journal of the Acoustical Society of America*, **137**, 2395, doi:10.1121/1.4920721.

Karnowski, J., M.-A. Yair, 2015: Classification of blue whale D calls and fin whale 40 Hz calls using deep learning. *Proc. the 7th international DCLDE workshop*, La Jolla, CA, Scripps Institution of Oceanography. [Available online at http://www.cetus.ucsd.edu/dclde/docs/pdfs/Monday/14-Karnowski.pdf.]

Madhusudhana, S. K., M. A. Roch, E. M. Oleson, M. S. Soldevilla, and J. A. Hildebrand, 2009: Blue whale B and D call classification using a frequency domain based robust contour extractor. *OCEANS 2009 - EUROPE*, Bremen, 2009, 1–7. doi: 10.1109/OCEANSE.2009.5278220.

Margolina, T., 2010: High frequency automatic recording package data summary report PS05, August 4, 2008–January 6, 2009, NPS Project Report NPS-OC-10-003, 40 pp.

Nieukirk, S. L., K. M. Stafford, D. K. Mellinger, R. P. Dziak, and C. G. Fox, 2004: Low-frequency whale and seismic airgun sounds recorded in the mid-Atlantic Ocean. *Journal of the Acoustical Society of America*, **115**, 1832–1843, doi:10.1121/1.1675816.

Oleson, E. M., J. A. Hildebrand, J. Calambokidis, G. Schorr, and E. Falcone, 2007a: 2006 progress report on acoustic and visual monitoring for Cetaceans along the outer Washington coast, NPS Project Report NPS-OC-07-003, 30 pp.

Oleson, E. M., J. Calambokidis, W. C. Burgess, M. A. McDonald, C. A. LeDuc, and J. A. Hildebrand, 2007b: Behavioral context of call production by eastern North Pacific blue whales. *Marine Ecology Progress Series*, **330**, 269–284, doi:10.3354/meps330269.

Oleson, E. M., S. M. Wiggins, and J. A. Hildebrand, 2007c: Temporal separation of blue whale call types on a southern California feeding ground. *Animal Behaviour*, **74**, 881–894, doi:10.1016/j.anbehav.2007.01.022.

Rocha, R. C., Jr., P. J. Clapham, and Y. V. Ivashchenko, 2014: Emptying the oceans: A summary of industrial whaling catches in the 20th century. *Marine Fisheries Review*, **76**, 37–48, doi:10.7755/MFR.76.4.3.

Scripps Institution of Oceanography, 2015: Dataset retrieval for the 2015 DCLDE workshop. Accessed 2 October 2016. [Available online at http://www.cetus.ucsd.edu/dclde/dataset.html.]

Scripps Institution of Oceanography, 2015: Scoring tools for the 2015 DCLDE workshop. Accessed 2 October 2016. [Available online at http://www.cetus.ucsd.edu/dclde/scoringTool.html]

Shamir, L., C. Yerby, R. Simpson, A. M. von Benda-Beckmann, P. Tyack, F. Samarra, P. Miller, and J. Wallin, 2014: Classification of large acoustic datasets using machine learning and crowdsourcing: application to whale calls. *Journal of the Acoustical Society of America,* **135**, 953–962, doi:10.1121/1.4861348.

Širović, A., 2011: Marine mammal demographics of the outer Washington coast during 2008–2009, NPS Project Report NPS-OC-11-004CR, 41 pp.

Širović, A., L. Williams, S. Kerosky, S. Wiggins, and J. Hildebrand, 2013: Temporal separation of two fin whale call types across the eastern North Pacific. *Mar. Biol.,* **160**, 47–57, doi:10.1007/s00227-012-2061-z.

Širović, A., J. A. Hildebrand, S. M. Wiggins, M. A. McDonald, S. E. Moore, and D. Thiele, 2004: Seasonality of blue and fin whale calls and the influence of sea ice in the Western Antarctic Peninsula. *Deep-Sea Research Part II,* **51**, 2327–2344, doi:10.1016/j.dsr2.2004.08.005.

Thode, A. M., G. L. D'Spain, and W. A. Kuperman, 2000: Matched-field processing, geoacoustic inversion, and source signature recovery of blue whale vocalizations. *The Journal of the Acoustical Society of America,* **107**, 1286–1300, doi:10.1121/1.428417.

Thompson, P. O., 1965: Marine biological sounds west of San Clemente Island: diurnal distributions and effects of ambient noise during July1963. Research report NEL/report 1290, 48 pp.

Thompson, P. O., L. T. Findley, and O. Vidal, 1992: 20-Hz pulses and other vocalizations of fin whales, Balaenoptera physalus, in the Gulf of California, Mexico. *Journal of the Acoustical Society of America,* **92**, 3051–3057, doi:10.1121/1.404201.

Watkins, W. A., 1982: Activities and underwater sounds of fin whales. *Deep Sea Research Part B. Oceanographic Literature Review,* **29**, 789, doi:10.1016/0198-0254(82)90294-1.

Wiggins, S. M., E. M. Oleson, M. A. McDonald, and J. A. Hildebrand, 2005: Blue whale (Balaenoptera musculus) diel call patterns offshore of Southern California. *Aquatic Mammals,* **31**, 161–168, doi:10.1578/AM.31.2.2005.161.

Wiggins, S. M., and J. A. Hildebrand, 2007: High-frequency Acoustic Recording Package (HARP) for broad-band, long-term marine mammal monitoring. *Symposium on Underwater Technology and Workshop on Scientific Use of Submarine Cables and Related Technologies*, 551–557, doi:10.1109/UT.2007.370760.

Watkins, W. A., M. A. Daher, G. M. Reppucci, J. E. George, D. L. Martin, N. A. DiMarzio, and D. P. Gannon, 2000: Seasonality and distribution of whale calls in the North Pacific. *Oceanography,* **13**, 62–67, doi:10.5670/oceanog.2000.54.

Yaser S. Abu-Mostafa, M. Magdon-Ismail, and H. Lin, 2012: *Learning from data.* AMLbook, 201 pp.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
    Ft. Belvoir, Virginia

2. Dudley Knox Library
    Naval Postgraduate School
    Monterey, California